STATISTICAL PROBLEMS IN USING MARKOV CHAIN TO REPRESENT DNA SEQUENCES AND THEIR APPLICATIONS

By

KIL-SUP LIM

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1998

ACKNOWLEDGEMENTS

As chairman of my committee, Dr. Mark C.K. Yang guided me throughout my dissertation. Without his encouragement and supporting, this work would never have been completed. I extend to him my sincere gratitude and will remeber him always. I would also like to thank Dr. Richard L. Scheaffer, Dr. Randy L. Carter, Susan P. McGorray, and Dr. Li-Min Fu for serving on my dissertation committee.

Dr. Eun-Sang Won, director of the Department for Force Development in Korea Institute for Defense Analyses, supported and encouraged me throughout my years at the University of Florida. I offer my sincere thanks to him.

I wish to express my special thanks to my family: my wife Yoonsook, for her love and patience; and to our sons, Jae-woong and Jae-min, for being a glorious joy to us. I am grateful to my parents and parents-in-law for their encouragement.

Finally, I would like to thank all my colleagues and friends for their assistance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO SYMBOLS

$A \propto B$: A is proportional to B.

$A \approx B$: A is approximately equal to B.

$A >> B$: $\frac{B}{A}$ is close to 0.

$A := B$: A is defined as B.

$A_n \sim B_n$: The ratio of A_n and B_n tends to unity.

$A_n \xrightarrow{\text{D}} B$: A_n converges in distribution to B.

$A_n \xrightarrow{\text{a.s.}} B$: A_n almost surely converges to B.

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

STATISTICAL PROBLEMS IN USING MARKOV CHAIN TO REPRESENT DNA SEQUENCES AND THEIR APPLICATIONS

By

Kil-Sup Lim

August 1998

Chairman: Mark C.K. Yang
Major Department: Statistics

With the advent of DNA sequencing techniques, researchers now have the opportunity to probe DNA sequences in search of stochastic or hidden structure. The potential benefits of finding the hidden structure are great as it may eventually guide biological modeling and experiments.

The purpose of this research is to provide the statistical tools to identify the hidden structure in DNA sequences in view of Markov chain frame and linguistic features.

The relationship between the order of the Markov chain for a DNA sequence and an amino acid sequence is identified by means of an expanded Markov chain technique. Also, a method based on the homogeneity property of a Markov chain representation of exon and intron is developed to identify the exon regions in DNA sequences. Improvement is observed when compared with a method based on correlation.

The linguistic feature of DNA sequences is investigated by means of Zipf's law. A method to compare two estimates of Zipf's coefficients is developed using the Pareto type variation model for the underlying distribution of word frequencies. It is observed, by applying the developed procedure to the given data sets, that the linguistic

features in exon regions are significantly different from those in intron regions and that the linguistic features in the intron regions can be explained by a lower order Markov chain than in the exon regions.

CHAPTER 1
INTRODUCTION

1.1 Motivation of Research

The heredity of genetic information and the evolution of genes are two of the most challenging problems facing evolutionary and molecular biologists. The principles by which nature produces the genetic information and the evolution process of DNA sequences are still not well understood.

It is believed that most of the principles or information are stored in DNA sequences. Experimental biologists have attempted to find the local level (deterministic) rules in DNA sequences by experimentation. On the other hand, theoretically oriented scientists have endeavored to discover whether *hidden structure* globally exists in DNA sequences.

With the advent of DNA sequencing techniques, researchers have had the opportunity to probe DNA for the hidden structure. As a result, several important issues have been raised, including Markov chain properties (Cuticchia et al., 1992), long-range correlations (Peng et al., 1992), and the connection to linguistics (Mantegna et al., 1995).

The benefits of finding the hidden structure would be great in the sense that they may eventually guide biological modeling and experiments. The importance of searching the stochastic rules in DNA sequences motivated this dissertation.

1.2 Scope of Research

Deoxyribonucleic acid (DNA) carries the information necessary for the development, maintenance, and reproduction of all organisms, from bacteria to humans. The stored information in DNA is first transcribed into ribonucleic acid (RNA). Then, after appropriate processing and splicing, RNA is turned into messenger RNA (mRNA), which is then translated into protein. The proteins in turn catalyze the many reactions taking place in the cell and serve as structural components of cellular organelles and membranes.

DNA sequences can be represented by a sequence of codes consisting of four letters (or bases, or nucleotides) denoting chemical products: adenine (A), thymine (T), cytosine (C), and guanine (G). These genetic codes are organized into words, which are called *codon*s, of three letters each. Codons are further grouped into *gene*s or parts of genes known as *exon*s, which are used to code proteins. DNA sequences also contain the non-exon parts, called *intron*s, sitting between exons. That is, introns split the gene into pieces. The function of the introns is mostly unknown. In RNA, a sister compound to DNA, uracil (U) substitutes the thymine (T) in DNA. To form mRNA, the exons in a DNA sequence become an uninterrupted sequence, whereas the introns are spliced out and discarded. Proteins are chains of *amino acid*s which contain 20 varieties. At the end of the gene, a stop signal stops the transcription of the protein. Figure 1.1 summarizes the flow of information from gene to protein. This flow process is the focus of the current research.

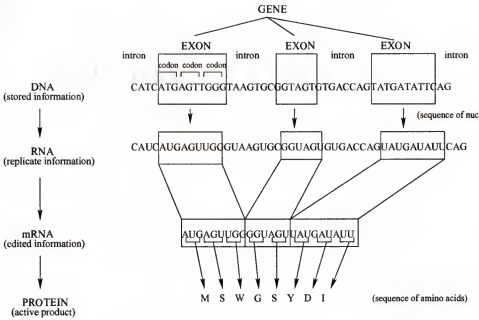

Figure 1.1. Information flow from gene to protein.

1.3 Purpose of Research

Two points of view are maintained in this dissertation. One is a Markovian sense on the consecutive structure of the DNA sequence and the amino acid sequence. The other is a linguistic view on overall usage of DNA blocks. This view is based on the resemblance of DNA sequences to natural language.

This research intends to provide statistical tools to search for the hidden structure in DNA sequences by utilizing the Markov chain properties and linguistic features. In particular, the relationship between a DNA sequence and an amino acid sequence is investigated in the context of Markov chains and then this relationship is used to develop a new algorithm to distinguish exons from introns. The linguistic features in a DNA sequence are measured by Zipf's law. Statistical properties of Zipf's law are inspected and the relationship between the Markov chain and linguistic features are also identified.

This dissertation is divided into four chapters. Chapter 2 deals with Markov chain properties of DNA sequences. The linguistic features in DNA sequences and their relation to Markov chain properties are discussed in Chapter 3. Both Chapters 2 and 3 contain background, literature review, summary and discussion. Chapter 4 summarizes the results and presents a brief discussion of future research issues.

CHAPTER 2
MARKOV CHAIN PROPERTIES IN DNA SEQUENCES

2.1 Background

2.1.1 Markov Chain in DNA Sequences

The whole DNA codes (represented as base pairs (bp)) for one living organism, called genome, are huge. For example, a human genome contains approximately 3×10^9 bases. To investigate the genome, one needs to determine how these DNA codes are arranged. At the local level, the codes in DNA sequences are deterministically translated into specific amino acids for a protein, but at the global level, it is known that the codes are not randomly arranged in a statistical sense (e.g., Arnold et al., 1987; Cuticchia et al., 1992). Thus, the mechanism by which DNA sequences are produced suggests that analysis of such sequences within the framework of stochastic process might be profitable. For example, the natural frequencies of many important oligonucleotide segments of DNA sequences may be described by a Markov chain model.

It is also known that the amino acids in protein sequences are not randomly distributed in global sense (Chou and Fasman, 1987). Therefore, it is interesting to investigate how the Markovian property of the codes in DNA sequences passes to the amino acid sequence in protein and what is the passage difference between exon and intron regions. The results of such an investigation should help in separating the

exon and intron regions in the genome. A very important area of DNA research is to distinguish exons from introns.

2.1.2 Expanded and Lumped Processes in DNA Sequences

To describe a DNA sequence, let $\{X_t; t = 1, 2, .., l_X\}$ be a discrete stochastic process, where X_t represents the value of a nucleotide at a position t with state space $\Omega_X = \{A, T, C, G\}$ and l_X is the length of the DNA sequence. Then, as seen in Figure 1.1, the corresponding codon sequence and amino acid sequence can be defined by $\{Y_t; t = 1, 2, \cdots, l_Y\}$ and $\{Z_t; t = 1, 2, \cdots, l_Z\}$, respectively, such that

$$Y_t = (X_{3t-2}, X_{3t-1}, X_{3t}),\ t = 1, 2, \cdots, l_Y \tag{2.1}$$

$$Z_t = f(Y_t),\ t = 1, 2, \cdots, l_Z (= l_Y), \tag{2.2}$$

where f is a function given by the rule (known) of converting 64 triplet codons to corresponding 21 amino acids (including stop signal). Table 2.1 shows the converting rule from codons to amino acids. In the table, we note that each T (thymine) in DNA is replaced by a U (uracil) in RNA.

For example, if

$$\{X_t\} = \{A, T, G, A, G, T, T, G, G, \cdots\},$$

then $\{Y_t\} = \{(ATG), (AGT), (TGG), \cdots\}$ and $\{Z_t\} = \{M, S, W, \cdots\}$, where M, S, and W denote the amino acids, Met (methionine), Ser (serine), and Trp (tryptophan), respectively.

The sequence $\{Y_t\}$, which represents codons, forms a larger process derived from $\{X_t\}$ in the sense that the state space of $\{Y_t\}$, $\Omega_Y = \{(AAA), (AAT), (AAC), \cdots\}$ with 64 elements, is enlarged from Ω_X with 4 states. We call $\{Y_t\}$ an *expanded process* of $\{X_t\}$. On the other hand, the sequence of amino acids $\{Z_t\}$ is an aggregated process from $\{Y_t\}$. The state space of $\{Z_t\}$, $\Omega_Z = \{M, S, W, \cdots\}$, with 21 elements is a many to one combination of Ω_Y. The process $\{Z_t\}$ is called a *lumped process* of $\{Y_t\}$.

Table 2.1. Coverting rule from codon to amino acid (3 character abbreviation).

	2nd				
	T	C	A	G	
1st					3rd
T	Phe(F)	Ser(S)	Tyr(Y)	Cys(C)	T
	Phe(F)	Ser(S)	Tyr(Y)	Cys(C)	C
	Leu(L)	Ser(S)	TC	TC	A
	Leu(L)	Ser(S)	TC	Trp(W)	G
C	Leu(L)	Pro(P)	His(H)	Arg(R)	T
	Leu(L)	Pro(P)	His(H)	Arg(R)	C
	Leu(L)	Pro(P)	Gln(Q)	Arg(R)	A
	Leu(L)	Pro(P)	Gln(Q)	Arg(R)	G
A	Ile(I)	Thr(T)	Asn(N)	Ser(S)	T
	Ile(I)	Thr(T)	Asn(N)	Ser(S)	C
	Ile(I)	Thr(T)	Lys(K)	Arg(R)	A
	Met(M)	Thr(T)	Lys(K)	Arg(R)	G
G	Val(V)	Ala(A)	Asp(D)	Gly(G)	T
	Val(V)	Ala(A)	Asp(D)	Gly(G)	C
	Val(V)	Ala(A)	Glu(E)	Gly(G)	A
	Val(V)	Ala(A)	Glu(E)	Gly(G)	G

(a) Source (modefied): Waterman, 1995.
(b) (): one character abbreviation.
(c) TC: termination signal.
(d) T in DNA substitutes U in RNA.

Therefore, it can be considered that the information stored in a gene is transmitted to the sequence of amino acids through expanding and lumping. This chapter deals with the information flow through the expanded and lumped processes when the DNA sequence is a Markov chain.

2.1.3 Purpose of Chapter 2

Even though the study of DNA sequences as Markov chains has received a great deal of attention in recent years, it is still just beginning. In particular, the information flow from DNA sequences to amino acid sequences has not been well investigated in the Markovian sense. Also, the difference in Markov chain properties between exon and intron regions is relatively unknown.

The main goals of this chapter are as described below. First, a model selection method (especially, the order determination for a Markov chain) is discussed as a first step to represent the dependence structure in the DNA sequence. Second, the relationship between the DNA and corresponding amino acid sequence is investigated in the Markovian sense. This study is focused on the difference between exon and intron regions in view of the expanded and lumped processes. Finally, an algorithm to distinguish exon regions from intron regions is developed based on the Markov chain properties of DNA sequences.

Section 2 of this chapter reviews the statistical and biological literature, including the representation of DNA sequences by Markov chains, model selection procedures for Markov chains, the expanded and lumped process, and an existing algorithm for detection of exon regions. Section 3 is devoted to the order selection procedures of the Markov chain. Section 4 discusses the Markovian relation between DNA and amino acid sequences in view of the expanded and lumped processes. An exon-intron identification principle is provided in Section 5. In Section 6, the algorithm

for detecting exon regions based on previous findings is developed. Section 7 contains summary and discussion.

2.2 Literature Review

2.2.1 Markov Chain Representation of DNA Sequences

Several studies have used Markov chain structure to represent DNA sequences.

Bishop et al. (1983) used a first-order Markov chain to investigate restriction sites, which are specific sequences of bases along a DNA sequence recognized by restriction enzymes. For example, the enzyme AluI recognizes the sequence $AGCT$. They estimated the transition matrix for the state space $\Omega = \{A, T, C, G, AG, AGC, AGCT\}$. Almagor (1983) also considered $\{X_t\}$ as a first-order Markov chain for predicting the natural frequencies of trinucleotides.

Blaisdell (1985) examined the option of using first-, second-, and third-order Markov chains to describe $\{X_t\}$ for the eukaryotic nuclear DNA sequences and used a χ^2 statistic to measure the difference between the observed and predicted transition matrices within a multiple hypothesis test framework. He found that most of the sequences require at least a second order for their representation and some sequences need the third-order chains. A third-order Markov chain has been used to predict oligonucleotide frequencies in E.coli genomes (Phillips et al., 1987), in Yeast DNA sequences (Arnold et al., 1987) and in Drosophila melanogaster DNA sequences (Cuticchia et al., 1992).

Prum et al. (1995) used a first-order Markov chain to find oligonucleotides which show unexpected frequencies in DNA sequences. They derived asymptotically standard normal statistics to test the extremeness of the frequency of an oligonucleotide under the assumption that the DNA sequence was a first-order Markov chain.

However, all of the previous studies, except Blaisdell's (1985) work, did not attempt to select the best Markov chain to fit DNA sequences. Most of the studies only showed that the frequencies of oligonuleotides can be predicted more accurately by using the first-order (Almagor, 1983) or the third-order (Phillips et al., 1987; Arnold et al., 1987; Cuticchia et al., 1992) Markov chain. For example, Cuticchia et al. (1992) used the following equation: $P(GGATCC) = P(C|CTA) \times P(C|TAG) \times P(T|AGG) \times P(AGG)$, to predict the frequency of $GGATCC$ segments. Blaisdell (1985) executed the multiple hypothesis test procedure by consecutively using the asymptotic χ^2 distribution, based on Anderson and Goodman (1957).

The next section reviews two popular procedures for choosing the order of a Markov chain, which have been suggested as alternative methods to the multiple hypothesis test technique.

2.2.2 Model Selection of Markov Chain

Among several methods proposed for estimating the order of a Markov chain, widely used procedures are the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These procedures attempt to balance the criteria of better fitting and parsimony, as measured by the number of unknown parameters.

Akaike (1974) has recommended a model selection criterion for cases when there are several competing models. The criterion is to select the model to minimize $-2\log(\text{maximum likelihood})+2$ (number of estimable parameters). The AIC procedure, based on information theoretic concepts, was suggested as an alternative to the multiple hypothesis test technique. Tong (1975) has applied the AIC procedure

to the Markov chain. He assumed that the order of the Markov chain is known to be less than some fixed upper bound $m(\geq 1)$. Then, the AIC estimator of the order of the Markov chain, $\hat{k}_{AIC}$, is chosen such that

$$AIC(\hat{k}_{AIC}) = min_{0\leq k<m} AIC(k), \tag{2.3}$$

where, for a given sample of l_X observations of Markov chain $X_1, \cdots, X_{l_X}$ with s states,

$$AIC(k) = -2\log\frac{M_k(X_1, \cdots, X_{l_X})}{M_m(X_1, \cdots, X_{l_X})} - 2(s^m - s^k)(s-1).$$

Here, $M_k(X_1, \cdots, X_{l_X})$ denotes the maximum likelihood function for the k^{th}-order transition probability given by $\Pi_{x_1,\cdots,x_{k+1}}(\frac{f_{x_1,\cdots,x_{k+1}}}{f_{x_1,\cdots,x_k.}})^{f_{x_1,\cdots,x_{k+1}}}$, where $f_{x_1,\cdots,x_{k+1}}$ denotes the number of transitions from states $x_1 \to x_2 \to, \cdots, \to x_{k+1}$ that occur in the sample and $f_{x_1,\cdots,x_k.} = \sum_{x_{k+1}} f_{x_1,\cdots,x_{k+1}}$.

Schwarz (1978) has pointed out that the maximum likelihood principle used by Akaike invariably leads to the possibility of overestimation. As an alternative to the AIC procedure, he has suggested the BIC procedure to select a model to minimize $-2\log$(maximum likelihood)+(number of estimable parameters)$\log l_X$.

Katz (1981) defined the BIC estimator ($\hat{k}_{BIC}$) of the order of a Markov chain, in a form similar to Tong's AIC estimator. That is,

$$BIC(\hat{k}_{BIC}) = min_{0\leq k<m} BIC(k), \tag{2.4}$$

with

$$BIC(k) = -2\log\frac{M_k(X_1, \cdots, X_{l_X})}{M_m(X_1, \cdots, X_{l_X})} - (s^m - s^k)(s-1)\log l_X.$$

He showed that the BIC procedure provides a consistent estimator of the Markov chain order. Also, he demonstrated that $\hat{k}_{AIC}$ has a positive asymptotic probability of overestimating the true order. For a finite sample, he compared the performance

of two procedures by using simulation. Two sets of samples (meteorological data) with binary (0 or 1) states were used by controlling a first order. The result based on the first data set showed that the BIC procedure performs well relative to the AIC procedure over the whole range of sample sizes, whereas the simulation based on the second data set demonstrated that the AIC procedure provides better estimates than the BIC procedure for sample sizes less than 1,500.

2.2.3 Expanded and Lumped Process

An *expanded process* is defined as described below. Let $\{X_t; t = 1, 2, ..l_X\}$ be a stochastic process taking values in a state space Ω_X with s_X elements. A new process $\{Y_t; t = 1, 2, \cdots, l_Y\}$ formed by combining n consecutive X_i's is called the expanded process of $\{X_t\}$ with length n. That is,

$$Y_t = (X_{nt}, X_{nt+1}, \cdots, X_{nt+(n-1)}). \tag{2.5}$$

Equation (2.5) is a more generalized form of equation (2.1). We note that the state space Ω_Y of $\{Y_t\}$ has $(s_X)^n$ elements, which are an enlarged set from Ω_X. Moreover, if $\{Y_t\}$ is a Markov chain, we call it an expanded Markov chain of $\{X_t\}$ with length n.

Previous studies on expanded the Markov chains are not plentiful. Kemeny and Snell (1976) discussed the special case of equation (2.5) with $s_X = 2$ and $n = 2$ for a first-order Markov chain. They were interested in examining the properties of the transition matrix of an expanded chain.

A *lumped process* is made by aggregating the states of an original process. Let $\{Y_t; t = 1, 2, \cdots, l_Y\}$ be a stochastic process taking values in a state space Ω_Y with s_Y elements. Assume that $\Omega_Z = \{G(1), \cdots, G(s_Z)\}$ is a partition of Ω_Y, where $s_Z < s_Y$ and $G(i)'s$ $(i = 1, \cdots, s_Z)$ represent disjoint subsets of Ω_Y with $\cup_{i=1}^{s_Z} G(i) = \Omega_Y$. Then, a new process $\{Z_t; t = 1, \cdots, l_Z\}$ is formed as follows. The outcome of the jth experiment in the new process is the set $G(i)$ that contains the outcome of the jth

step in the original process $\{Y_t\}$. Thus, $l_Y = l_Z$ and $s_Y > s_Z$. This new process $\{Z_t\}$ is called a lumped process with state space Ω_Z. Equation (2.2) is an example of a lumped process.

A great deal of attention has been placed on the conditions under which a lumped process forms a Markov chain. A original process $\{Y_t\}$ (usually, a Markov chain) is said to be lumpable with respect to a partition $\Omega_Z = \{G(1), \cdots, G(s_Z)\}$ if the resulting lumped process $\{Z_t\}$ is a Markov chain satisfying some conditions. There are two kinds of lumpablity, strong and weak, according to the conditions.

Strong lumpability is defined as follows (Kemeny and Snell, 1976): A Markov chain $\{Y_t\}$ with state space $\Omega_Y = \{1, 2, \cdots, s_Y\}$ is strongly lumpable with respect to a partition $\Omega_Z = \{G(1), \cdots, G(s_Z)\}$ of Ω_Y if, for every initial probability vector of $\{Y_t\}$, $P_Y{}^0$, the resulting chain $\{Z_t\}$ (with the state space of dimension $1 < s_Z < s_Y$) is a Markov chain and the transition probabilities of $\{Z_t\}$, $p_Z(i,j) = P[Z_{t+1} = j | Z_t = i]$, are invariant to the choice of P_Y^0.

Kemeny and Snell (1976) also provided a theorem for strong lumpability; if, for a partition $\Omega_Z = \{G(1), \cdots, G(s_Z)\}$ of Ω_Y, the transition probabilities $p_Y(k, G(j))$ are defined such that

$$p_Y(k, G(j)) = \sum_{l \in G(j)} p_Y(k, l) \quad \text{for} \quad k \in G(i)$$

then a necessary and sufficient condition for a Markov chain to be strongly lumpable with respect to a partition Ω_Z is that for every pair of $G(i)$ and $G(j)$, the $p_Y(k, G(j))$ have the same value for every $k \in G(i)$. These common values form the transition probabilities, $p_Z(i,j)$, for the lumped chain.

Thomas and Barr (1977) provided a practical test procedure to check for strong lumpability. They developed the maximum likelihood estimators of the unknown transition matrix of $\{Y_t\}$, P_Y, under the null hypothesis that $\{Y_t\}$ is strongly lumpable

with respect to a given partition of Ω_Y. Then, they adopted the usual χ^2 test procedure to test the null hypothesis. In addition, Thomas (1977) suggested a computational procedure to test strong lumpability in the case that the transition matrix P_Y is known.

By the lumping, we generally obtained a more manageable chain at the sacrifice of obtaining less precise information. Lindqvist (1978) measured the loss of information incurred by observing a strongly lumped Markov chain instead of the original Markov chain.

In contrast with strong lumpability, a process is defined to be *weakly lumpable* with respect to a partition Ω_Z of Ω_Y if at least one initial probability vector,${P_Y}^0$, leads to a Markov chain (Kemeny and Snell, 1976).

Rubino and Sericola (1989) characterized the sufficient condition for weak lumpability by finding a set of all initial probability vectors that lead to a lumped Markov chain. In a followup paper (1991), they suggested a method to find the set of all inital probability vectors which permit a Markov chain to be weakly lumpable given the transition matrix and partition of the state space.

Recently, Norberg (1997) suggested a procedure to check weak lumpability, when transition matrix is known, by finding the distribution of sojourn time that a Markov chain spends in a subset $G(i)$ of the lumped state space $\Omega_Z = \{G(1), \cdots, G(s_Z)\}$.

2.2.4. Existing Exon-Intron Identification Method

Identifying the exon regions of an uncharacterized genomic DNA sequence is one of the difficult tasks in DNA sequence analysis. During the last few years, several complex programs for detecting exon regions have been developed. Most of these programs make use of multiple lines of evidence (i.e., measures) from which to draw conclusions about the position of exon regions.

In this subsection, the typical measures used to detect exon regions are first introduced and then the widely used programs identifying genes or exon regions are discussed.

Measures to Detect Exon Regions

The measures to distinguish exon regions from intron regions generally take two forms: search by content and search by signal. Content statistics measure the bulk properties over a length of the sequence, whereas signal-based methods look for short functional sequence elements which indicate the existence of an exon. Typically, these are boundaries around exon and intron regions.

The traditional approach to identify the functional sites is to search for similarity of the query sequence, which is a candidate for a functional site, to reference sequences. The reference represents the common features of a group of known sequences with similar functions. One modern method of representing the reference is a matrix in which the occurrence of each nucleotide at each position in the vicinity of the functional sites is calculated based on the known sequences. For example, each element in the matrix $f(b, i)$ represents the number of nucleotide b found at position i. Then, we compare the query sequence with the matrix and make a decision on the function of the query sequence.

There are several variants of the reference matrix and the method of comparing. However, a detailed discussion is omitted because the new measure found by this research can be classified as a content statistic rather than a signal-based measure. A comprehensive review of various signal-based measures was done by Snyder (1994).

The most typical content statistic is probably the *Codon usage* measure, which was developed in several studies (e.g., Staden and McLachlan, 1982). The codon usage measure is effective for detecting exon regions because most organisms have a biased (unequal) usage of codons and most proteins have a biased usage of amino acids.

Therefore, we can identify exon regions by comparing the frequencies of codons in an unknown DNA sequence with a reference set, which contains the codon frequencies of an already known DNA sequence.

There are other content statistics which are not related to codon usage. Periodicities in nucleotide usage occur differently according to the characteristics of the sequence type. Exon regions tend to display a three-base repeat $XNNX$ (where X is a specific base and N can be any base), while intron regions display a two-base repeat XNX (Arques and Michel, 1987). These properties can be measured by *Autocorrelation.*

A third type of content statistic is *Base compositional bias*, which measures the asymmetry of the base composition of the three codon positions (Fickett and Tung, 1992). This is based on the fact that codons in exon regions are often of the form RNY (R:purin (A or G), Y:pyrimidine (C or T), N:either) or WWS (W:weak hydrogen bonding (A or T), S: strong hydrogen bonding (C or G)). Thus, exon regions are detected by measuring the difference between the unknown DNA sequence and the pattern RNYRNY $\cdots$ RNY, or WWSWWS $\cdots$ WWS.

Another interesting type of content statistics is *Long-range correlation* which was recently discussed by several studies (e.g., Peng et al., 1992). They observed that intron regions of DNA sequences possess long-range power-law correlations, whereas exon regions typically do not. Ossadnik et al. (1994) suggested an algorithm, which is called CSF (Coding Sequence Finder), to measure the long-range correlation and to identify the exon regions in DNA sequences. They quantified the correlation properties of a DNA sequence by using a random walk model: The walker steps *up* $[U(i) = +1]$ if a pyrimidine (C or T) occurs at position i along the DNA chain, whereas the walker steps *down* $[U(i) = -1]$ if a purine (A or G) occurs at position i. The displacement of the walker after g steps, $y(g)$, is defined as $y(g) = \sum_{i=1}^{g} U(i)$. Here, the following two procedures are executed to investigate the correlation property of the

DNA sequence: (1) Divide the entire sequence of length l into l/w non-overlapping boxes, each containing w nucleotides, and define $\hat{y}_w(g)$ (local trend) in each box to be the ordinate of the least square fit of the walk displacements ($y(g)$'s). (2) Calculate $z_w(g) = |y(g) - \hat{y}_w(g)|$ in each box of size w and determine $F_d^2(w) = \frac{1}{l}\sum_{g=1}^{l} z_w^2(g)$ over all the boxes in the whole sequence. Then, the exon regions are predicted by observing the behavior of the estimate $\hat{\alpha}$ from the least square fit (by log-log transformation) of

$$F_d(w) \propto w^{\alpha}.$$

Here, Ossadnik et al. (1994) used the previous finding by Peng et al. (1992, 1994) that $\alpha \approx 0.5$ for exon regions, whrereas α is substantially larger than 0.5 for intron regions.

Programs to Combine Multiple Pieces of Evidence

There are two basic approaches to apply evidence to identify exon regions or genes (Snyder and Stormo, 1997): rule-based methods (GeneID, GeneModeler) and neural network methods (GeneParser, GRAIL2).

In the rule-based approach, criteria are applied serially to identify possible exons and then rank them or eliminate them from consideration. The neural network methods produce summary statistics that are applied in parallel and weighed according to their importance. Currently, GeneModeler supports the analysis of sequences from Homosapiens, Drosophila melanogaster, C. elegans, and Dicotyledonous and Monocotyledonous plants, while GeneID, GeneParser, and GRAIL2 were originally developed for use on mammalian (especially, human) genome.

GRAIL2 is probably the most widely used program for exon identification. It is available in three forms. The original version of GRAIL (Uberbacher and Mural, 1991), used via an e-mail server, provides a table of neural network outputs with predicted exon regions and a prediction quality. In this version, start and stop codons

are not found. The newly available GRAIL2 (Xu et al., 1994) uses the same basic GRAIL algorithm but further attempts to identify complete exon regions by using several signal-based measures. There are four primary steps to exon recognition in GRAIL2: (1) generation of the initial candidate pool of exons, (2) elimination of highly improbable candidates, (3) evaluation of remaining candidates with neural networks, and (4) clustering of scored candidates for final prediction. GRAIL2 reports likely exon regions on the forward and reverse strand in DNA sequence and makes a final table of results similar to the original GRAIL now with precise exon boundaries.

XGRAIL, which is the newest version, provides a graphical user interface as well as additional analysis functionalities.

When GRAIL2 was tested on 109 DNA sequences, it recognized 697 of 746 exon regions (93.5% of total true exon regions) and predicted 93 false exon regions (12% of total prediction). In the correctly predicted exons, 62% of the exons were found perfectly (with both edges correct) and 93% with at least one edge correct. Apparently, there is still room for improvement.

Another feature of GRAIL2 is its species-sensitivity. This is because the neutral network training procedure is sensitive to the training set used for optimizing the weights of the different measures. This implies that the parameters used in GRAIL2 need to be adjusted for other organisms, especially non-mammalian DNA sequences.

2.3 Order Selection of Markov Chains for DNA Sequences

As reviewed in Section 2, Katz (1981) compared the BIC and AIC procedures which are the most widely used order selection methods for Markov chain models. He showed that the BIC procedure provides a consistent estimator for the order. Also, he demonstrated, by simulations, that the BIC procedure performs well relative to

the AIC procedure in the case of two-state first-order Markov chains. However, the four-state DNA chain has not been considered.

To investigate the discrepancy between the BIC estimate($\hat{k}_{BIC}$) and the AIC estimate($\hat{k}_{AIC}$) when they are used to represent DNA sequences as Markov chains, 10 DNA sequences are chosen from the data set previously used by Mantegna et al. (1994) as an illustrative example. Table 2.2 shows the estimated order when equations (2.3) and (2.4) are applied to each DNA sequence. In this example, the BIC procedure indicates a second-order dependence for most sequences, while a third or fourth order is estimated by the AIC procedure. This result agrees with what Schwartz (1978) has pointed out; $\hat{k}_{AIC}$ tends to be greater than $\hat{k}_{BIC}$.

Table 2.2. BIC and AIC estimates[a] of the Markov chain order for some DNA sequences.

Organism		DNA sequence (locus)	Length (bp)	Estimates	
				BIC	AIC
I. eukaryotes					
	1. mammal	HUMTCRADCV	97634	2	4
		MMBGCXD	55856	2	3
	2. invertebrates	CELTWIMUSC	54962	2	3
		CEF59B2	43782	1	3
II. eukaryotic virsuses					
		ASFV55KB	55098	2	3
		IH1CG	134226	3	4
III. prokaryotes					
		BSGENR	97015	2	3
		ECOUW87	96484	3	4
IV. bacteriophages					
		LAMCG	48502	2	3
		MLCGA	52297	2	4

(a) Upper bound of order: 9.

Even though the BIC procedure yields a consistent estimator while AIC has an asymptotic probability of overestimation, typical DNA sequences may not be long enough to apply the asymptotic properties (usually, $100 - 300,000$ *bp*). So, it is important to investigate how $\hat{k}_{BIC}$ and $\hat{k}_{AIC}$ behave in the range of DNA sequence lengths. Analytical expressions for the exact distributions of $\hat{k}_{BIC}$ and $\hat{k}_{AIC}$ are not available, however. Moreover, they would probably be too complicated to be very useful (Katz, 1981). Therefore, simulation studies are commonly used to compare the performance of these procedures.

Considering that a Markov chain representation of DNA sequences 4 states are needed and that the estimated orders in the previous example (Table 2.2) range from first to fourth. It is of interest to investigate the order detection capabilities of BIC and AIC procedures when four-state chains of varing orders are considered. Two-state chains are also considered for comparison with Katz's (1981) work, in which he only dealt with a two-state chain controlled by a first order. Here, a simulation study is composed of two cases as follows: First, chains with two states are generated for first, second and third order based on the transition probabilities of Manchester weather data (Gates and Tong, 1976). Second, the first-, second-, and third-order chains with 4 states are generated by using the transition probabilities of a DNA sequence, SCCHRIII, with length $315,339$ *bp*.

Table 2.3 gives the ratio of correct detection, underestimation and overestimation for sample lengths ranging from $l = 50$ to $l = 128,000$ based on 100 replications when the BIC and AIC procedures are applied to two-state chains. The results show that agreement with the asymptotic properties is quite close. In particular, for two-state first-order chains, the BIC procedure performs satisfactorily relative to the AIC procedure. This result agrees with Katz's first simulation study. However, for second- and third-order chains, the AIC procedure provides better estimates for a relatively short sample length ($l \leq 8,000$ for second-order chains, $l \leq 16,000$ for third-order

chains). Additionally, it should be noted that, for a large sample, the AIC procedure gives the correct detection ratio fluctuated at some level, which tends to be increased as the true order is increased: roughly 0.65 for a first order, 0.80 for a second order, and 0.92 for a third order. Table 2.4 shows the detection ratios ranging from $l = 200$

Table 2.3. Performance of $\hat{k}^a_{BIC}$ and $\hat{k}^a_{AIC}$ for first-, second-, and third-order Markov chains with 2 states based on the Manchester weather data.

True order	Length (bp)	BIC			AIC		
		Correct detection	Under[b] estimat.	Over[c] estimat.	Correct detection	Under[b] estimat.	Over[c] estimat.
1	50	0.60	0.38	0.02	0.55	0.10	0.35
	100	0.92	0.06	0.02	0.62	0.01	0.37
	200	0.99	0.00	0.01	0.60	0.00	0.40
	400	1.00	0.00	0.00	0.67	0.00	0.32
	800	1.00	0.00	0.00	0.63	0.00	0.37
	1000	1.00	0.00	0.00	0.64	0.00	0.35
	128000	1.00	0.00	0.00	0.66	0.00	0.33
2	100	0.01	0.99	0.00	0.18	0.65	0.17
	1000	0.05	0.95	0.00	0.56	0.28	0.16
	2000	0.12	0.88	0.00	0.71	0.13	0.15
	4000	0.42	0.58	0.00	0.81	0.02	0.17
	8000	0.84	0.16	0.00	0.85	0.00	0.15
	16000	1.00	0.00	0.00	0.84	0.00	0.15
	64000	1.00	0.00	0.00	0.84	0.00	0.17
	128000	1.00	0.00	0.00	0.81	0.00	0.20
3	200	0.00	1.00	0.00	0.12	0.84	0.04
	800	0.00	1.00	0.00	0.18	0.78	0.04
	1000	0.01	0.99	0.00	0.53	0.39	0.08
	4000	0.19	0.81	0.00	0.89	0.02	0.11
	8000	0.35	0.65	0.00	0.91	0.01	0.08
	16000	0.84	0.16	0.00	0.92	0.00	0.08
	32000	1.00	0.00	0.00	0.94	0.00	0.06
	128000	1.00	0.00	0.00	0.93	0.00	0.07

(a) Upper bound of order:6.
(b) Underestimation.
(c) Overestimation.

to $l = 300,000$ based on 100 replications in the case of DNA sequences with 4 states. For moderately large samples, $\hat{k}_{AIC}$ nearly always makes the correct choice of the true order, whereas $\hat{k}_{BIC}$ provides correct detection for quite long chains. Additionally, the AIC procedure always gives better estimates than BIC in this table.

Based on this simulation study, the BIC procedure needs a much larger sample size to attain good asymptotic properties as the number of states and the true order of the chain are increased. However, the AIC procedure tends to provide good estimates for high order chains with a large number of states, even though there exists a probability of overestimation for low order chains with a small number of states (e.g., first-order chains with 2 states). Considering DNA sequences are frequently represented by high order (second, third, fourth) four-state Markov chains of finite length, the AIC procedure is a more adequate method to estimate the order of Markov chain representing DNA sequences. If the AIC procedure is adopted as the estimation method, the results in Table 2.2 (third orders are detected in most of DNA sequences) agree with previous studies (Blaisdell, 1985; Arnold et al., 1987; Phillips et al., 1987; Cuticchia et al., 1992), even though the data sequences and the estimation methods were different.

2.4 Relation between DNA and Amino Acid Sequences in a Markovian Sense

When the nucleotides in exon regions are used to code proteins, three consecutive nucleotides of DNA sequence (i.e., a codon) corresponds to one amino acid in the protein sequence. In symbolic representations such as equations (2.1) and (2.2), the codon sequence $\{Y_t; t = 1, 2, \cdots, l_Y\}$ can be considered as an expanded process of the

Table 2.4. Performance of $\hat{k}^a_{BIC}$ and $\hat{k}^a_{AIC}$ for first-, second-, and third-order Markov chains with 4 states based on the DNA sequence SCCHRIII.

True order	Length (bp)	BIC			AIC		
		Correct detection	Under[b] estimat.	Over[c] estimat.	Correct detection	Under[b] estimat.	Over[c] estimat.
1	400	0.00	1.00	0.00	0.38	0.62	0.00
	1000	0.00	1.00	0.00	0.77	0.23	0.00
	2000	0.00	1.00	0.00	0.97	0.02	0.00
	4000	0.00	1.00	0.00	1.00	0.00	0.00
	9000	0.03	0.97	0.00	1.00	0.00	0.00
	10000	0.27	0.73	0.00	1.00	0.00	0.00
	12000	0.66	0.34	0.00	1.00	0.00	0.00
	14000	0.99	0.01	0.00	1.00	0.00	0.00
	15000	1.00	0.00	0.00	1.00	0.00	0.00
2	1000	0.00	1.00	0.00	0.01	0.99	0.00
	4000	0.00	1.00	0.00	0.20	0.80	0.00
	8000	0.00	1.00	0.00	0.80	0.20	0.00
	16000	0.00	1.00	0.00	1.00	0.00	0.00
	60000	0.18	0.82	0.00	1.00	0.00	0.00
	70000	0.68	0.32	0.00	1.00	0.00	0.00
	80000	0.99	0.01	0.00	1.00	0.00	0.00
	90000	1.00	0.00	0.00	1.00	0.00	0.00
3	16000	0.00	1.00	0.00	0.23	0.77	0.00
	32000	0.00	1.00	0.00	0.99	0.01	0.00
	64000	0.00	1.00	0.00	1.00	0.00	0.00
	200000	0.02	0.98	0.00	1.00	0.00	0.00
	220000	0.17	0.83	0.00	1.00	0.00	0.00
	240000	0.77	0.23	0.00	1.00	0.00	0.00
	260000	0.99	0.01	0.00	1.00	0.00	0.00
	300000	1.00	0.00	0.00	1.00	0.00	0.00

(a) Upper bound of order:6.
(b) Underestimation.
(c) Overestimation.

corresponding DNA sequence $\{X_t; t = 1, 2, \cdots, l_X\}$, while the amino acid sequence $\{Z_t; t = 1, 2, \cdots, l_Z\}$ is a lumped process of the codon sequence $\{Y_t\}$.

Section 2.4.1 discusses the relationship between $\{X_t\}$ and $\{Y_t\}$ in view of expanded processes, but under a more general definition. The relationship between $\{X_t\}$ and $\{Z_t\}$ is given in Section 2.4.2.

2.4.1 Properties of Expanded Markov Chains

Definition 2.1 Let $\{X_t; t = 1, 2, \cdots, l_X\}$ be a discrete time stochastic process taking values in a countable state space Ω_X. A new process $\{Y_t; t = 1, 2, \cdots, l_Y\}$ is called an *expanded process* of $\{X_t\}$ with length n ($n \geq 2$), degree of overlapping v ($0 \leq v \leq n-1$), and starting point i ($0 \leq i \leq n-v-1$), if

$$Y_{(i)t} = (X_{(n-v)t+i}, X_{(n-v)t+i+1}, \cdots, X_{(n-v)t+i+n-1}) \tag{2.6}$$

where n, v and i are non-negative integers. Here, for all i, the state space of $\{Y_{(i)t}\}$ is the set of all possible combinations of the states of $\{X_t\}$ according to length n. Moreover, if $\{Y_{(i)t}\}$ is a Markov chain, it is called an *expanded Markov chain* of $\{X_t\}$.

For example, if $n = 3$ and $v = 1$, then

$$\{Y_{(0)t}\} = \{\cdots, (X_4, X_5, X_6), (X_6, X_7, X_8), (X_8, X_9, X_{10}), \cdots\}$$

and

$$\{Y_{(1)t}\} = \{\cdots, (X_5, X_6, X_7), (X_7, X_8, X_9), (X_9, X_{10}, X_{11}), \cdots\}.$$

Therefore, equation (2.1) is a special case of equation (2.6) when $n = 3, v = 0$ and $i = 0$, while equation (2.5) is another special case when $v = 0$ and $i = 0$. Also, if three different starting points are considered in the DNA sequence $\{X_t\}$, there are three codon sequences $(\{Y_{(0)t}\}, \{Y_{(1)t}\}, \{Y_{(2)t}\})$.

Kemeny and Snell (1976) discussed a special case of this problem with $n = 2$, $v = 0$ and $\Omega_X = \{0, 1\}$ for a first-order Markov chain. However, they were only

interested in examining the relationship between the transition matrices of $\{X_t\}$ and $\{Y_{(0)t}\}$ under the assumption that both of $\{X_t\}$ and $\{Y_{(0)t}\}$ are first-order Markov chains. Theorem 2.1 and Corollary 2.1 show the general relationships between $\{X_t\}$ and $\{Y_{(i)t}\}$.

Theorem 2.1 Let $\{X_t; t = 1, 2, \cdots, l_X\}$ be a stationary p^{th} $(p \geq 1)$ order Markov chain taking values in a countable state space and let the process $\{Y_{(i)t}; t = 1, 2, \cdots, l_Y\}$ be defined by equation (2.6). Then, $\{Y_{(i)t}\}$ is a q^{th} order Markov chain for any non-negative integers v and i such that $0 \leq v \leq n-1$ and $0 \leq i \leq n-v-1$, where q is given by

$$q = \lceil (p-v)/(n-v) \rceil, \tag{2.7}$$

and $\lceil x \rceil$ denotes the smallest positive integer greater than or equal to x.

Proof To simplify notation, let $\delta = n - v$. Then, for any non-negative integers v and i given in the theroem,

$$\begin{aligned}
& P[Y_{(i)t}|Y_{(i)t-1}, Y_{(i)t-2}, \cdots, Y_{(i)0}] \\
= \; & P[X_{\delta t+i+(n-1)}, X_{\delta t+i+(n-2)}, \cdots, X_{\delta t+i}|X_{\delta(t-1)+i+(n-1)}, \cdots, X_{\delta(t-1)+i} \\
& \quad , X_{\delta(t-2)+i+(n-1)}, \cdots, X_{\delta(t-2)+i}, \cdots, X_0] \\
= \; & P[X_{\delta t+i+(n-1)}|X_{\delta t+i+(n-2)}, \cdots, X_{\delta t+i}, X_{\delta(t-1)+i+(n-1)}, \cdots, X_{\delta(t-1)+i} \\
& \quad , X_{\delta(t-2)+i+(n-1)}, \cdots, X_{\delta(t-2)+i}, \cdots, X_0] \\
& \times P[X_{\delta t+i+(n-2)}|X_{\delta t+i+(n-3)}, \cdots, X_{\delta t+i}, X_{\delta(t-1)+i+(n-1)}, \cdots, X_{\delta(t-1)+i} \\
& \quad , X_{\delta(t-2)+i+(n-1)}, \cdots, X_{\delta(t-2)+i}, \cdots, X_0] \\
& \cdots \\
& \times P[X_{\delta t+i+(r-\delta)}|X_{\delta t+i+(n-\delta-1)}, \cdots, X_{\delta t+i}, X_{\delta(t-1)+i+(n-1)}, \cdots, X_{\delta(t-1)+i} \\
& \quad , X_{\delta(t-2)+i+(n-1)}, \cdots, X_{\delta(t-2)+i}, \cdots, X_0]
\end{aligned}$$

$$
\begin{aligned}
= \; & P[X_{\delta t+i+(n-1)}|X_{\delta t+i+(n-2)}, \cdots, X_{\delta t+i}, X_{\delta(t-1)+i+(n-1)}, \cdots, X_{\delta(t-1)+i}, \\
& \quad \cdots, X_{\delta(t-q)+i+(n-1)}, \cdots, X_{\delta(t-q)+i}] \\
& \times P[X_{\delta t+i+(n-2)}|X_{\delta t+i+(n-3)}, \cdots, X_{\delta t+i}, X_{\delta(t-1)+i+(n-1)}, \cdots, X_{\delta(t-1)+i}, \\
& \quad \cdots, X_{\delta(t-q)+i+(n-1)}, \cdots, X_{\delta(t-q)+i}] \\
& \cdots \\
& \times P[X_{\delta t+i+(n-\delta)}|X_{\delta t+i+(n-\delta-1)}, \cdots, X_{\delta t+i}, X_{\delta(t-1)+i+(n-1)}, \cdots, X_{\delta(t-1)+i}, \\
& \quad \cdots, X_{\delta(t-q)+i+(n-1)}, \cdots, X_{\delta(t-q)+i}],
\end{aligned}
$$

where $q(q \geq 1)$ is the smallest integer satisfying the following inequality.

$$\delta(t-q) \leq \delta t + (n-\delta) - p.$$

Then, since $\{X_t\}$ is a stationary Markov chain, the index i for starting points can be dropped. Thus, for any non-negative integer i such that $0 \leq i \leq n-v-1$, we have that

$$
\begin{aligned}
& P[Y_{(i)t}|Y_{(i)t-1}, Y_{(i)t-2}, \cdots, Y_{(i)0}] \\
= \; & \frac{P[X_{\delta t+(n-1)}, \cdots, X_{\delta t}, X_{\delta(t-1)+(n-1)}, \cdots, X_{\delta(t-1)}, \cdots, X_{\delta(t-q)+(n-1)}, \cdots, X_{\delta(t-q)}]}{P[X_{\delta(t-1)+(n-1)}, \cdots, X_{\delta(t-1)}, .., X_{\delta(t-q)+(n-1)}, \cdots, X_{\delta(t-q)}]} \\
= \; & P[Y_t|Y_{t-1}, \cdots, Y_{t-q}].
\end{aligned}
$$

Thus, $\{Y_t\}$ is a q^{th}-order Markov chain, where q is given by equation (2.7) □

Corollary 2.1 If $\{X_t; t = 1, 2, \cdots, l_X\}$ is a stationary Markov chain taking values in a countable state space, the resulting expanded processes according to different starting points, $\{Y_{(0)t}\}, \{Y_{(1)t}\}, \cdots, \{Y_{(n-v-1)t}\}$, form homogeneous Markov chains in the sense that they share a common transition matrix.

Proof It is obvious from the last step in the proof of Theroem 2.1. □

As mentioned earlier, codon sequences with 64 states are non-overlapping ($v = 0$) expansions with length (n) 3 from the original DNA sequences with 4 states. Therefore, three codon sequences are converted from a DNA sequence according to three different starting points (i). Here, Corollary 2.1 implies that the three codon sequences form homogeneous Markov chains in the sense of having the same transition matrices, provided the DNA sequence is a stationary Markov chain.

It is also possible to go back in the opposite direction to Theorem 2.1. Theorem 2.2 discusses the properties of an original process when its expanded process is given.

Theorem 2.2 For $i = 0, 1, \cdots, n-v-1$, let $\{Y_{(i)t}; t = 1, 2, \cdots, l_Y\}$ be a q^{th} $(q \geq 1)$ order expanded Markov chain from the process $\{X_t; t = 1, 2, \cdots, l_X\}$ as defined by equation (2.6). If $\{X_t\}$ is a stationary Markov chain taking values in a countable state space, then the order, p, of $\{X_t\}$ is determined by

$$(n-v)(q-1)+1 \leq p \leq (n-v)q, \text{ for all } i = 0, 1, \cdots, n-v-1, \qquad (2.8)$$

where n, v, and the state spaces of $\{Y_{(i)t}\}$ and $\{X_t\}$ are given in Definition 2.1.

Proof By using the stationarity of $\{X_t\}$, the index i of $\{X_t\}$ and $\{Y_{(i)t}\}$ for starting points in equation (2.6) can be dropped. Again, letting $\delta = n - v$,

$$\begin{aligned}
& P[X_{\delta(t+1)}|X_{\delta t+n-1}, \cdots, X_{\delta t}, X_{\delta(t-1)+n-1}, \cdots, X_{\delta(t-1)}, \cdots, X_0] \\
= & \frac{\sum_{X_{\delta(t+1)+n-1}, \ldots, X_{\delta(t+1)+1}} P[X_{\delta(t+1)+n-1}, \ldots, X_{\delta(t+1)+1}, X_{\delta(t+1)}, X_{\delta t+n-1}, \ldots, X_0]}{P[X_{\delta t+n-1}, \ldots, X_{\delta t}, \ldots, X_0]} \\
= & \frac{\sum_{X_{\delta(t+1)+n-1}, \ldots, X_{\delta(t+1)+1}} P[Y_{t+1} = (X_{\delta(t+1)+n-1}, \ldots, X_{\delta(t+1)+1}, X_{\delta(t+1)},}{P[Y_t = (X_{\delta t+n-1}, \ldots, X_{\delta t}), \ldots,} \\
& \frac{Y_t = (X_{\delta t+n-1}, \ldots, X_{\delta t}), \ldots, Y_{t-q+1} = (X_{\delta(t-q+1)+n-1}, \ldots, X_{\delta(t-q+1)}),}{Y_{t-q+1} = (X_{\delta(t-q+1)+n-1}, \ldots, X_{\delta(t-q+1)}),} \\
& \frac{X_{\delta(t-q)+n-1}, \ldots, X_0]}{X_{\delta(t-q)+n-1}, \ldots, X_0]}
\end{aligned}$$

$$= \sum_{X_{\delta(t+1)+n-1},\ldots,X_{\delta(t+1)+1}} P[Y_{t+1} = (X_{\delta(t+1)+n-1}, \ldots, X_{\delta(t+1)+1}, X_{\delta(t+1)}$$

$$|Y_t = (X_{\delta t+n-1}, \ldots, X_{\delta t}), \ldots, Y_{t-q+1} = (X_{\delta(t-q+1)+n-1}, \ldots, X_{\delta(t-q+1)})]$$

$$= P[X_{\delta(t+1)}|X_{\delta t+n-1}, \ldots, X_{\delta t}, X_{\delta(t-1)+n-1}, \ldots, X_{\delta(t-1)}, \ldots, X_{\delta(t-q+1)}],$$

where $\sum_{X_t}$ denotes that summation over all possible values of variable X_t. Note that the number of the conditional terms of the above equation is given by $(n-v)q$. But, if $\{Y_t\}$ does not have the full strength of q^{th} order dependence, that is, if all components (X_t's) in Y_{t-q+1} of the above expressions are not needed to explain Y_{t+1} even though $\{Y_t, t \geq 0\}$ is a q^{th}-order chain, then the number of conditional terms can be reduced to $(n-v)(q-1)+1$. Thus, $\{X_t\}$ is a p^{th}-order Markov chain, where p is given by equation (2.8). Also, due to the stationary property of $\{X_t\}$, the above argument holds for arbitrary starting points for $\{X_t\}$. □

2.4.2 Relationship between DNA and Amino Acid Sequences

Again using the index i to denote starting points in expanding a DNA sequence, the amino acid sequence $\{Z_{(i)t}\}$ can be expressed as the lumped process of the codon sequence $\{Y_{(i)t}\}$. That is,

$$Z_{(i)t} = f(Y_{(i)t}), \quad t = 1, 2, \cdots, l_Z (= l_Y), \tag{2.9}$$

where f is a function given by the rule shown in Table 2.1. Here, the state space of $\{Z_{(i)t}\}$ (21 elements) is a partition of the state space of $\{Y_{(i)t}\}$ (64 elements). We can consider lumpability, which was reviewed in Section 2. However, strong lumpability may be too restrictive to apply to complex DNA sequences, even though Thomas and Barr (1977) suggested a practical test procedure. Unfortunately, a test method for conveniently checking weak lumpability is not available.

Therefore, instead of lumpability, we consider the estimation problem of the Markov chain models for representing of the original process, its expanded process,

Table 2.5. AIC estimates of the Markov chain order for codon and amino acid sequences corresponding to several DNA sequences.

DNA sequence	Length(bp)	Sequence	Estimated order by AIC (i: starting point) $i=0$	$i=1$	$i=2$
HUMTCRADCV	97634	DNA[a]	4^d		
		Codon[b]	1	1	1
		Amino acid[c]	1	1	1
IH1CG	134226	DNA[a]	4^d		
		Codon[b]	1	1	1
		Amino acid[c]	1	1	1
BSGENR	97015	DNA[a]	3^d		
		Codon[b]	1	1	1
		Amino acid[c]	1	1	1

(a) Original process.
(b) Expanded process of a.
(c) Lumped process of b.
(d) Not relevant to starting points.

and its lumped process. Table 2.5 demonstrates the estimated order of codon and amino acid sequences corresponding to DNA sequences selected from Table 2.2. All amino acid sequences lumped from corresponding codon sequences are estimated to be first-order Markov chains according to each starting points $i = 0, 1, 2$.

From Table 2.5, we note that codon sequences expanded from some DNA sequences, HUMTCRADCV and IH1CG, are estimated as first order, even though second-order chains are expected in view of Theorem 2.1. This may be because the sequences are not long enough to reject the first-order hypothesis.

The relationship between the DNA and amino acid sequences can be examined using Corollary 2.1. This corollary implies that, if the DNA sequence $(\{X_t\})$ lacks the stationarity, then the corresponding codon sequences $(\{Y_{(i)t}\}, i = 0, 1, 2)$ may not form the same Markov chains when the starting points are different. In this case, the corresponding amino acid sequences $(\{Z_{(i)t}\}, i = 0, 1, 2)$ lumped from $\{Y_{(i)t}\}$ may not form homogeneous Markov chains, either.

The above observation provides a useful tool to examine DNA sequences. More specifically, if stationary parts and non-stationary parts are mixed in a DNA sequence, then we can distinguish between the two kinds of parts by testing the homegeneity of three Markov chains for codons (or amino acids) converted from the corresponding part of DNA sequence. However, using the codon sequence (64 states) is impractical because there are too many parameters to be estimated in the transition matrix. Therefore, the converted amino acid sequence (21 states) is preferred for testing homogeneity.

2.5 Principle for Exon-Intron Identification

In this section, results from previous sections are applied to several DNA sequences to test the homogeneity of three Markov chains according to three different starting points when the exon and intron parts in a DNA sequence are converted into an amino acid sequence. The difference in homogeneity between exon and intron regions, if present, can serve as a measure for exon-intron identification.

For application, three data sets, *DATA-WHOLE*, *DATA-EXON*, and *DATA-INTRON*, are used. *DATA-WHOLE* includes 9 DNA sequences, which were taken from the GeneBank (May, 1997), of three different groups: Yeast (Saccharomyces cerevisiae, 3

sequences), Chloroplast (3 sequences), and C. elegans (3 sequences). Table 2.6 shows the contents of *DATA-WHOLE*. For each of DNA sequences in *DATA-WHOLE*, exon regions were extracted and combined into one sequence of exon parts (forming *DATA-EXON*), while the remaining fragments composed the other sequence of intron parts (forming *DATA-INTRON*).

Table 2.6. Contents of the data set *DATA-WHOLE* (exon regions are known).

Sequence (locus)	Length (*bp*)	Ratio of exon region(%)	Number of exon region	Length of a exon region(average,*bp*)
1. Yeast				
SCCHRIII	315339	68.4	176	1381.9
YSCCHRVIN	270184	67.2	128	1476.5
SCCHRXVI	165536	71.7	82	1447.0
2. Chloroplast				
CHMPXX	121024	58.9	101	727.8
CHNTXX	155844	52.4	115	824.9
CHOSXX	134525	48.7	123	564.3
3. C. elegans				
CELC50C3	44733	48.1	66	326.0
CELF44E2	33651	42.2	42	338.4
CELTWIMUSC	54962	39.3	31	696.5

To test the homogeneity of the three sequences of amino acids, the likelihood ratio principle is used. Before testing, two propositional conditions should be checked: (1) the AIC procedure is used to determine the orders of the Markov chains for exon and intron sequences within DNA sequences (Table 2.7). Thus, we make sure that the sequences can be expressed by appropriate Markov chain models. (2) The sequences of amino acids converted from the exon and intron sequences of the DNA sequences can be also expressed by Markov chains by AIC procedure. They are either zeroth- or first-order Markov chains (Table 2.8). Some of the zeroth-order cases may be because

the data is not large enough to reject the zeroth order hypothesis. Since zeroth order is a special case of first order, we regard all sequences of amino acids generated from *DATA-EXON* and *DATA-INTRON* as first-order Markov chains.

Table 2.7. Estimates of the Markov chain order for DNA sequences by the AIC procedure[a].

Sequence (locus)	*DATA-WHOLE* (order)	*DATA-EXON* (order)	*DATA-INTRON* (order)
1. Yeast			
SCCHRIII	4	4	4
YSCCHRVIN	3	3	3
SCCHRXVI	3	3	3
2. Chloroplast			
CHMPXX	4	3	4
CHNTXX	4	4	4
CHOSXX	4	3	4
3. C. elegans			
CELC50C3	3	3	2
CELF44E2	3	3	2
CELTWIMUSC	3	4	3

(a) Upper bound of order:9.

For the three Markov chains of order 1 with 3 different starting points, we can test the null hypothesis that the chains are homogeneous, that is, that they share the same transition probability matrix. Let $P_i(k,j)$ and f_{ijk} be the transition probability and observed frequency from state j to state k in the i^{th} reading frame, respectively, where $i = 0, 1, 2$ and $j, k = 1, \cdots, 21$. Then, the null hypothesis that $P_i(k,j)$ is the same for all i and for all pairs of j and k can be tested by the statistic T, where

$$T = 2\sum_{i=1}^{3}\sum_{j=1}^{21}\sum_{k=1}^{21} f_{ijk}\log\frac{f_{ijk}}{f_{ij.}f_{.jk}}, \tag{2.10}$$

Table 2.8. Estimates of the Markov chain order by the AIC procedure[a] for amino acid sequences converted from DNA sequences according to starting points (i).

Sequence (Locus)	*DATA-EXON*				*DATA-INTRON*			
	Length (bp)	$i=0$	$i=1$	$i=2$	Length (bp)	$i=0$	$i=1$	$i=2$
1. Yeast								
SCCHRIII	215571	1	1	1	99768	1	1	1
YSCCHRVIN	181611	1	1	1	88573	1	1	1
SCCHRXVI	118658	1	1	1	46878	1	1	1
2. Chloroplast								
CHMPXX	71325	1	1	1	49699	1	1	1
CHNTXX	81663	1	1	1	74181	1	1	1
CHOSXX	65463	1	1	1	69062	1	1	1
3. C. elegans								
CELC50C3	21516	0	0	1	23217	0	0	0
CELF44E2	14211	0	0	0	19440	0	0	0
CELTWIMUSC	21591	1	1	1	33371	0	0	0

(a) Upper bound of the order:2.

$$with\ f_{ij.} = \sum_{k=1}^{21} f_{ijk},\ f_{jk} = \sum_{i=1}^{3} f_{ijk},\ f_{.j.} = \sum_{i=1}^{3}\sum_{k=1}^{21} f_{ijk}.$$

According to a basic theroem for likelihood ratio tests, T has an asymptotic χ^2 distribution with 840 degrees of freedom (Kullback et al., 1962).

Table 2.9 shows the result of χ^2 tests on *DATA-EXON* and *DATA-INTRON*. The three Markov chains of amino acids, according to three different starting points, in all exon sequences are significantly different, whereas 8 cases of 9 intron sequences show homogeneity when tested with significance level 0.05. All χ^2 statistics for exon sequences are greater than the corresponding statistics for intron sequences.

Table 2.9. χ^2 test of homogeneity among three Markov chains for amino acids (critical value=908.537 with significance level 0.05).

Sequence	*DATA-EXON*		*DATA-INTRON*	
(locus)	χ^2-statistic	p-value	χ^2-statistic	p-value
1. Yeast				
SCCHRIII	1290.3	0.0000	825.0	0.6378
YSCCHRVIN	2903.2	0.0000	884.0	0.1421
SCCHRXVI	5810.2	0.0000	828.6	0.6040
2. Chloroplast				
CHMPXX	4374.8	0.0000	893.6	0.0973
CHNTXX	1159.8	0.0000	829.2	0.5976
CHOSXX	1174.2	0.0000	865.1	0.2671
3. C. elegans				
CELC50C3	2392.3	0.0000	923.8	0.0229
CELF44E2	1562.1	0.0000	877.1	0.1817
CELTWIMU	5359.4	0.0000	848.7	0.4099

These results strongly imply that the intron regions in DNA sequences possess the stationary property in contrast with the exon regions. Consequently, the χ^2 statistic, T, can be used as a measure for detecting exon regions in DNA sequences.

2.6 Algorithm for Exon-Intron Identification

One of major problems facing researchers working with long genomic DNA sequences is the need for a rapid and accurate method for identifying exon regions. Currently, as reviewed in Section 2, a typical search for an exon region involves scanning the DNA sequence to detect a possible region which may contain the exon. The region is then searched for the existence of signal-based measures by using appropriate data sets. These methods are labor intensive and require considerable operator participation. In contrast, an ideal technique should not only be fast and accurate but also require minimal operator input.

Also, most of programs to detect exon regions are species-sensitive, because each uses a particular training set. For example, the famous GRAIL2, as pointed out in Section 2, performs a training procedure on the data set of human DNA sequences to optimize weights of the different sensor algorithms to initially detect the possible exon regions. However, because most sensor algorithms are species-sensitive, algorithm parameters need to be adjusted for other organisims. Therefore, an algorithm based on a more general principle across the entire phylogenetic spectrum would be desirable.

Recently, Ossadnik et al. (1994) has suggested such a tool (Coding Sequence Finder (CSF) algorithm) to identify exon regions based on power-law long-range correlation. They showed that 68%-74% of predictions correspond to exon regions when the CSF algorithm was applied to Yeast and Chloroplast genomes. Apparently, there is still room for improvement.

Since the existing methods are still far from perfection, new methods of gene identification should be sought. Hopefully, by combining new methods with the existing ones, the overall performance can be improved. The method to be discussed here is

an algorithm that can be applied to the entire phylogenetic spectrum. It uses Markov chain properties found in previous sections.

Algorithm for Markovian Exon Finder (MEF)

The new method is based on the observation that the existence of stationary Markov chain property in intron regions are not the same as in exon regions. In particular, the key concept of this algorithm is to use the fact that the sequences of 21 amino acids converted from exon regions form different Markov chains when the starting points of a reading frame change. A sliding reading box (*window*) along the DNA sequence is used to detect exon regions by continuously testing the homogeneity of the three Markov chains with three starting points. We call this method the *Markovian Exon Finder (MEF).*

The MEF algorithm calculates the χ^2 test statistic, T in equation (2.10), for the segment (window) of the DNA sequence. If the calculated T shows signifcance (e.g, the critical value is 908.537 with significance level 0.05), it means that the segment is more likely to be exon region. Details of the MEF algorithm are as follows:

Step 1. Select an appropriate window size (w). Even though smaller window size generally provide more accurate prediction for a short exon sequence, there is a limitation because of the sample size requirement in estimating first-order transition probabilities when there are 21 states. We found, by examining performance of the algorithm, that a suitable window size for the sequences in our data set would be 1,300-1,700 *bp*.

Step 2. Convert the nucleotide sequence in the window into three sequences of amino acids according to starting points 0, 1, and 2.

Step 3. Calculate T in equation (2.10) for the three sequences of amino acids according to equation and denote it by $T(x)$, where x is the mid-point of the window position in the DNA sequence.

Step 4. Slide the window along the DNA sequence and repeat Step 3. The sliding process may not be consecutive at every base in the DNA sequence. The next window starts after a given number of nucleotides, called the *jumping size* (m), to save computing time. Typically, a smaller jumping size produces higher sensitivity, lower specificity and requires more computing time (sensitivity and specificity will be defined later).

Step 5. Smooth out the resulting function $T(x)$ to remove statistical noise. A linearly weighted moving average with 5-9 points turns out to be an effective smoother. Let p be the number of points to be smoothed. The MEF algorithm, then, requires three parameters (w, m, p).

Step 6. Get the local maxima in the series of smoothed $T(x)$'s. If local maximum $T(x^*)$ is significant by the χ^2 test at the 0.05 level, then x^* is called the *predicted point.* This point is then claimed as one point in the exon region. We define the *predicted band* as $(x^* - 20,\ x^* + 20)$ *bp.* We use a band of width 40 *bp* because, if it is a part of the true exon region, then the probability of sequencing such a band matched by any part of a genome is approximately $(4^{41})^{-1} = 1.2 \times 10^{-24}$. Hence, the predicted band should uniquely determine the gene when it matches with a cDNA in a cDNA library.

Figure 2.1 shows the known exon regions and predicted points by applying step 1-6 to the well-known SCCHRIII sequence.

Evaluation of MEF Algorithm

To quantitatively characterize the accuracy of the MEF algorithm, we use the standard measures, *sensitivity* and *specificity,* to evaluate the performance of the predicted bands. The sensitivity is defined as *sensitivity*=P[*correct prediction* | *actual*

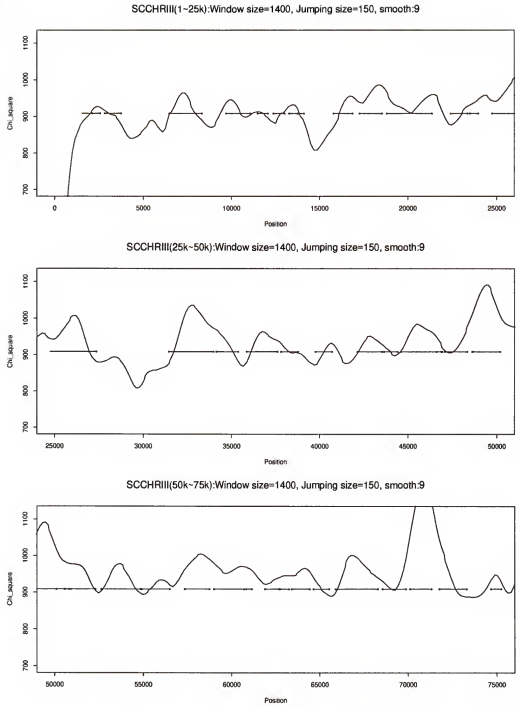

Figure 2.1. Analysis of a part of SCCHRIII by the MEF algorithm (window size=1400 *bp*, jumping size=150 *bp*, smoothed points=9).

(1) Considered region: 1-75,000 *bp* (see the horizontal axis).
(2) The horizontal line indicates known exon regions and curve shows the smoothed chi-square statistcs ($T(x)$). The predicted points are local maxima greater than the critical value for chi-square test (908.537).

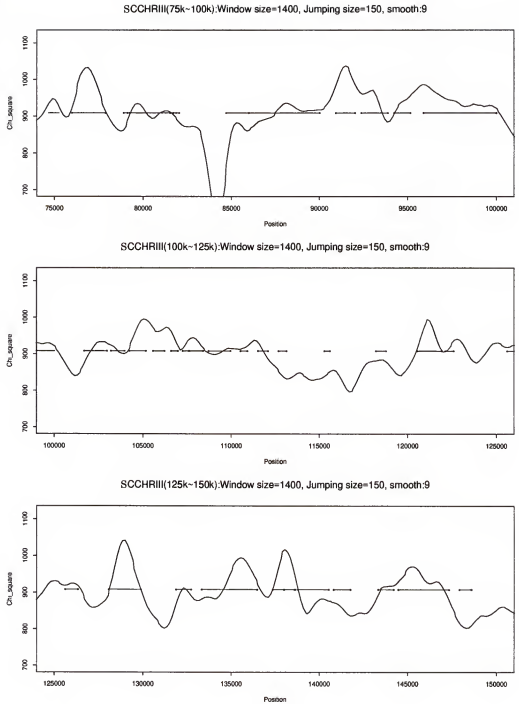

Figure 2.1-continued.
(1) Considered region: 75,000-150,000 *bp* (see the horizontal axis).
(2) The horizontal line indicates known exon regions and curve shows the smoothed chi-square statistcs ($T(x)$). The predicted points are local maxima greater than the critical value for chi-square test (908.537).

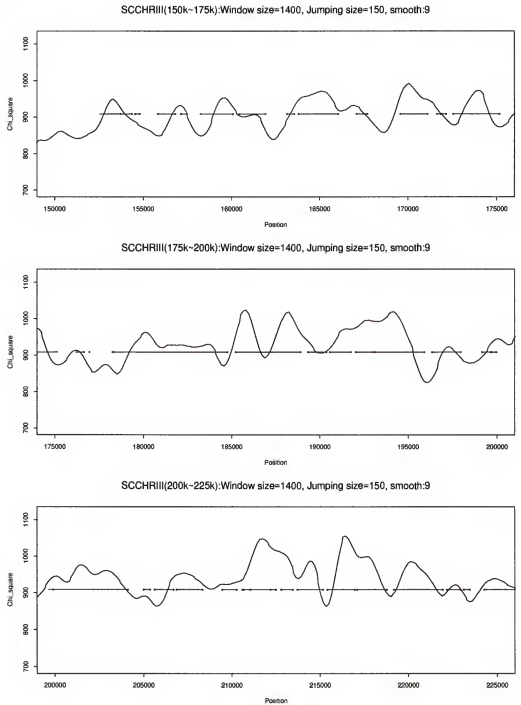

Figure 2.1-continued.
(1) Considered region: 150,000,000-225,000 *bp* (see the horizontal axis).
(2) The horizontal line indicates known exon regions and curve shows the smoothed chi-square statistcs ($T(x)$). The predicted points are local maxima greater than the critical value for chi-square test (908.537).

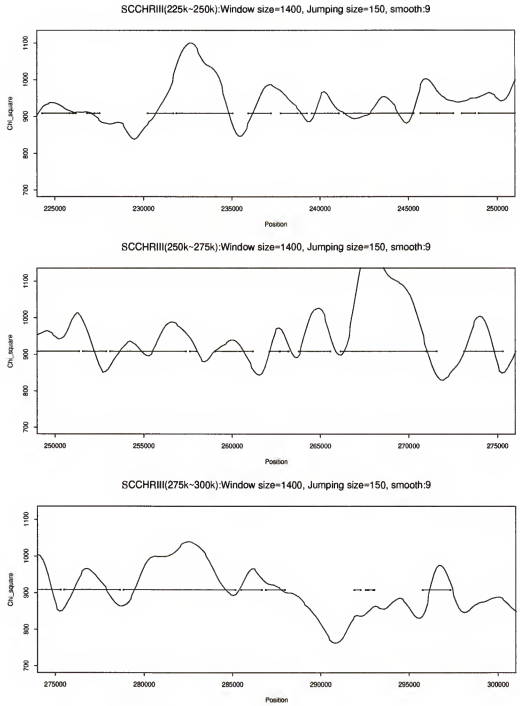

Figure 2.1-continued.
(1) Considered region: 225,000-300,000 *bp* (see the horizontal axis).
(2) The horizontal line indicates known exon regions and curve shows the smoothed chi-square statistcs ($T(x)$). The predicted points are local maxima greater than the critical value for chi-square test (908.537).

exon region], which is estimated by

$$sensitivity = \frac{number\ of\ actual\ exon\ regions\ correctly\ predicted}{number\ of actual\ exon\ regions}. \tag{2.11}$$

Also, the specificity is defined as *specificity*=P[*correct prediction* | *predicted exon region*], which is estimated by

$$specificity = \frac{number\ of correctly\ predicted\ regions}{number\ of\ predicted\ regions}. \tag{2.12}$$

Here, *correct prediction* means that the entire predicted band is included in the actual exon region. We note that the definition of the specificity in the literature may not be same as the equation (2.11) (e.g., Robert et al., 1982). However, in this research, the equation is used for emphasizing the accuracy of a prediction on the exon regions.

Sensitivity and specificity can not be increased simultaneously. They have the trade-off relationship similar to Type I and Type II errors in statistical hypothesis testing (Figure 2.2). If the purpose of exon region identification is to identify fragments that are surely in exon regions, then specificity is preferred. In general, sensitivity increases as the number of predicted regions increases.

The detailed results from *DATA-WHOLE* of known organisms are shown in Table 2.10 - 2.13 when a predicted band of width 40 *bp* was used with various algorithm parameters (w, m, p).

In this research, the parameter set is chosen according to following criteria: (1) Sensitivity and specificity are equally prefered, that is, consider *Sum= sensitivity+specificity.* (2) If the *Sums* are similiar, then the parameter set with the highest specificity is chosen. Applying the above criteria to Table 2.10 - 2.13 leads to a parameter setting $(w, m, p) = (1400, 5, 9)$. To measure the general performance of MEF with the chosen parameters, a new data set, which is called *DATA-NEW*, was formed from the GeneBank as follows: (1) 24 new sequences from the same organism as *DATA-WHOLE* (Yeast and C. elegans) and (2) 4 sequences from new organism (Helecobacteria pylori) which was just recently announced (August, 1997).

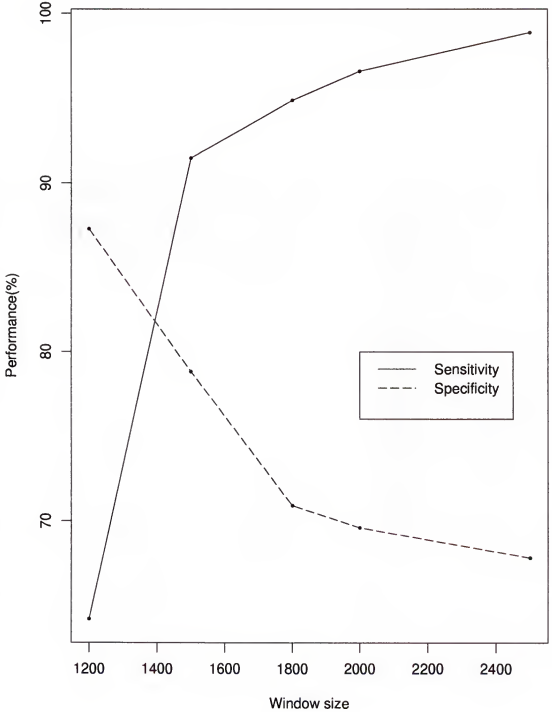

Figure 2.2. Relation between sensitivity and specificity.

The performance is measured on the predicted band (width=40 *bp*) along the whole sequence of SCCHRIII with jumping size=5, smoothed points=9.

Table 2.10. Accuracy of predicted band (width=40 *bp*) by MEF on DATA-WHOLE with parameter set (w, m, p)=(1300,m,9).

Jumping size (m)	Sequence (locus)	Length (*bp*)	Sensitivity (%,a)	Specificity (%,b)	Sum(a+b)
m=5	SCCHRIII	315339	76.14	84.36	160.50
	YSCCHRVIN	270184	78.91	82.61	161.52
	SCCHRXVIN	165536	74.39	86.28	160.67
	CHMPXX	121024	19.80	90.11	109.91
	CHNTXX	155844	38.26	80.67	118.93
	CHOSXX	134525	21.14	84.44	105.58
	CELC50C3	44733	25.76	68.22	93.98
	CELF44E2	33651	26.19	80.73	106.92
	CELTWIMUSC	54962	22.58	79.23	101.81
m=50	SCCHRIII	315339	56.82	82.50	139.32
	YSCCHRVIN	270184	64.84	83.62	148.46
	SCCHRXVIN	165536	59.76	85.46	145.22
	CHMPXX	121024	16.83	92.00	108.83
	CHNTXX	155844	23.48	75.00	98.48
	CHOSXX	134525	10.57	70.37	80.94
	CELC50C3	44733	10.61	56.25	66.86
	CELF44E2	33651	14.29	76.92	91.21
	CELTWIMUSC	54962	19.36	75.00	94.36
m=100	SCCHRIII	315339	48.30	87.29	135.59
	YSCCHRVIN	270184	57.03	86.79	143.82
	SCCHRXVIN	165536	52.44	93.44	145.88
	CHMPXX	121024	14.85	93.75	108.60
	CHNTXX	155844	19.13	76.67	95.80
	CHOSXX	134525	10.57	86.67	97.24
	CELC50C3	44733	6.06	57.14	63.20
	CELF44E2	33651	14.29	87.50	101.79
	CELTWIMUSC	54962	16.13	83.33	99.46

Table 2.11. Accuracy of predicted band (width=40 *bp*) by MEF on DATA-WHOLE with parameter set (w, m, p)=(1400,m,9).

Jumping size (m)	Sequence (locus)	Length (*bp*)	Sensitivity (%,a)	Specificity (%,b)	Sum(a+b)
m=5	SCCHRIII	315339	84.66	81.37	166.03
	YSCCHRVIN	270184	84.38	77.95	162.33
	SCCHRXVIN	165536	85.37	81.70	167.07
	CHMPXX	121024	21.78	80.92	102.70
	CHNTXX	155844	46.09	78.63	124.72
	CHOSXX	134525	30.89	76.90	107.79
	CELC50C3	44733	34.85	65.96	100.81
	CELF44E2	33651	33.33	77.45	110.78
	CELTWIMUSC	54962	29.03	67.76	96.79
m=50	SCCHRIII	315339	65.91	80.82	146.73
	YSCCHRVIN	270184	70.31	80.47	150.78
	SCCHRXVIN	165536	67.07	82.17	149.24
	CHMPXX	121024	17.82	81.82	99.64
	CHNTXX	155844	25.22	69.36	94.58
	CHOSXX	134525	18.70	76.92	95.62
	CELC50C3	44733	10.61	42.11	52.72
	CELF44E2	33651	16.67	76.47	93.14
	CELTWIMUSC	54962	19.36	56.52	75.88
m=100	SCCHRIII	315339	52.84	84.29	137.13
	YSCCHRVIN	270184	57.03	80.00	137.03
	SCCHRXVIN	165536	65.85	92.21	158.06
	CHMPXX	121024	16.83	85.71	102.55
	CHNTXX	155844	23.48	75.00	98.48
	CHOSXX	134525	13.82	78.26	92.08
	CELC50C3	44733	9.09	50.00	59.09
	CELF44E2	33651	14.29	87.50	101.79
	CELTWIMUSC	54962	19.36	85.71	105.07

Table 2.12. Accuracy of predicted band (width=40 *bp*) by MEF on DATA-WHOLE with parameter set (w, m, p)=(1500,m,9).

Jumping size (m)	Sequence (locus)	Length (*bp*)	Sensitivity (%,a)	Specificity (%,b)	Sum(a+b)
m=5	SCCHRIII	315339	91.48	78.81	170.29
	YSCCHRVIN	270184	86.72	75.24	161.96
	SCCHRXVIN	165536	87.81	77.50	165.31
	CHMPXX	121024	24.75	77.07	101.82
	CHNTXX	155844	52.17	74.48	126.65
	CHOSXX	134525	40.65	69.57	110.22
	CELC50C3	44733	42.42	60.62	103.05
	CELF44E2	33651	30.95	70.09	101.04
	CELTWIMUSC	54962	41.94	59.71	101.65
m=50	SCCHRIII	315339	71.02	81.11	152.13
	YSCCHRVIN	270184	71.88	77.25	149.13
	SCCHRXVIN	165536	74.39	82.52	156.91
	CHMPXX	121024	21.78	76.32	98.10
	CHNTXX	155844	27.83	75.76	103.59
	CHOSXX	134525	22.76	73.21	95.97
	CELC50C3	44733	12.12	36.36	48.48
	CELF44E2	33651	16.67	80.00	96.67
	CELTWIMUSC	54962	22.58	53.57	76.15
m=100	SCCHRIII	315339	53.98	83.92	137.90
	YSCCHRVIN	270184	60.94	80.33	141.27
	SCCHRXVIN	165536	65.85	84.34	150.19
	CHMPXX	121024	17.82	81.82	99.64
	CHNTXX	155844	23.48	73.91	97.39
	CHOSXX	134525	16.26	66.67	82.93
	CELC50C3	44733	9.09	53.85	62.94
	CELF44E2	33651	14.29	75.00	89.29
	CELTWIMUSC	54962	22.58	72.22	94.80

Table 2.13. Accuracy of predicted band (width=40 *bp*) by MEF on DATA-WHOLE with parameter set (w, m, p)=(1600,m,9).

Jumping size (m)	Sequence (locus)	Length (*bp*)	Sensitivity (%,a)	Specificity (%,b)	Sum(a+b)
m=5	SCCHRIII	315339	92.05	75.77	167.82
	YSCCHRVIN	270184	90.63	72.59	163.22
	SCCHRXVIN	165536	92.68	76.51	169.19
	CHMPXX	121024	28.71	75.92	104.63
	CHNTXX	155844	60.00	70.08	130.08
	CHOSXX	134525	45.53	63.67	109.20
	CELC50C3	44733	43.94	60.89	104.83
	CELF44E2	33651	30.95	68.18	99.13
	CELTWIMUSC	54962	45.16	59.42	104.58
m=50	SCCHRIII	315339	72.16	75.00	147.16
	YSCCHRVIN	270184	73.44	76.75	150.19
	SCCHRXVIN	165536	75.61	80.00	155.61
	CHMPXX	121024	22.77	76.92	99.69
	CHNTXX	155844	31.30	72.50	103.80
	CHOSXX	134525	25.20	69.84	95.04
	CELC50C3	44733	16.67	45.83	62.50
	CELF44E2	33651	14.29	66.67	80.96
	CELTWIMUSC	54962	22.58	48.39	70.97
m=100	SCCHRIII	315339	55.68	81.38	137.06
	YSCCHRVIN	270184	59.38	77.69	137.07
	SCCHRXVIN	165536	62.20	80.25	142.45
	CHMPXX	121024	17.82	73.07	90.89
	CHNTXX	155844	24.35	65.96	90.31
	CHOSXX	134525	20.33	66.67	86.99
	CELC50C3	44733	6.06	30.77	36.83
	CELF44E2	33651	11.91	62.50	74.41
	CELTWIMUSC	54962	22.58	63.16	85.74

Table 2.14. Accuracy of predicted band (width=40 *bp*) by MEF on *DATA-NEW* with determinded parameter set (w, m, p)=(1400,5,9).

Sequence (locus)	Length (*bp*)	Sensitivity (%,a)	Specificity (%,b)	Sum(a+b)
Yeast				
SC4357	75317	75.00	86.72	161.72
SC4987	21330	80.00	89.69	169.69
SC8339	27559	80.00	84.66	164.66
SC8520X	41200	90.00	83.39	173.39
SC9402	38990	100.00	75.42	175.42
SC9150	43504	75.00	83.50	158.50
SC9168	42793	79.17	74.46	153.63
SC9959	40397	82.61	83.67	166.28
SCD8035	69023	79.49	84.65	164.14
SCD9461	78500	76.47	81.33	157.80
SCD9717	72119	81.82	82.89	164.71
SCE6592	42576	73.91	75.88	149.79
SCE8199	36872	77.27	78.05	155.32
SCE9537	66030	86.11	79.92	166.03
SCE9781	68302	75.76	79.44	155.20
SCE9871	62643	71.88	81.95	153.83
SCU12980	103687	91.49	78.80	170.29
YSCH8179	44100	74.07	75.33	149.40
YSCH9196	55069	88.46	80.80	169.26
YSCLTESPO	29410	100.00	84.03	184.03
C. elegans				
CELC32E8	43726	19.70	55.39	75.09
CELF12B6	29268	35.00	30.38	65.38
CELT12F5	32707	42.86	51.28	94.14
CELF53G12	41553	30.16	38.54	68.70
Helicobacteria pylori				
HPAE000539	10631	72.73	89.06	161.79
HPAE000557	16693	73.33	91.67	165.00
HPAE000608	11074	40.00	100.00	140.00
HPAE000651	10028	63.64	93.75	157.39

Table 2.14 demonstrates the performance of MEF when applied to *DATA-NEW* with a parameter set at (1400,5,9). The performance on the new sequences in Yeast and C. elegans appears to be similar to the previous result (Table 2.11). The prediction accuracy on the sequences in the new organism (Helecobacteria pylori) is slightly worse than in the case of Yeast. This is because the new organism contains shorter exon regions (roughly, in the range of 800-1,000 *bp*) than Yeast (roughly, in the range of 1,000-1,700 *bp*).

For a comparison with other algorithms, we first consider Ossadnik et al.'s (1994) CSF algorithm, which was developed for a similar purpose as MEF. The CSF algorithm showed the specificity (sensitivity was not known) of predicted points as follows: 78% for SCCHRIII with w=800 *bp*, 74% for CHMPXX with w=1,000 *bp*, 72% for CHNTXX with w=1,200 *bp*, and 68% for CHOSXX with w=2,200 *bp*. On the other hand, the accuracy of predicted points (not predicted band) by the MEF algorithm turned out to be 83% ,82%, 82%, 78% for SCCHRIII, CHMPXX, CHNTXX, CHOSXX, respectively, with the determined parameter set (w=1,400 *bp*, m=5 *bp*, p=9). Therefore, substantial improvement (in view of specificity) has been made when compared with CSF algorithm.

For a comparision with the widely used GRAIL2, DNA sequences of length less than 100,000 *bp* were selected from *DATA-WHOLE* because of the input limitation of GRAIL2. By using the e-mail version of GRAIL2, the accuracy was measured on the sequences. The definitions in (2.11) and (2.12) were applied to calculate the accuracy of the predicted band of width 40 *bp* around the mid-range of exon region predicted by GRAIL2 and MEF. Table 2.15 shows the performance of GRAIL2 and MEF on *DATA-WHOLE*. Also, Table 2.16 compares the performance of GRAIL2 with MEF on *DATA-NEW*. Here, the determined parameter set $(1400, 5, 9)$ was used over all sequences when MEF was applied. Combing Table 2.15 and 2.16, we note that the

MEF algorithm showed better prediction accuracy in the sense of *Sum* (=sensitivity+specificity) for Yeast, Chloroplast and Helicobacteria, while GRAIL2 and MEF gave similar results for C. elegans. This means that, even though GRAIL2 combines several complex rules to make a prediction and has been used to detect exon regions in mammalian (especially, human) DNA with good predictive power, the identification can be improved by using a simple method based on the Markov chain property.

Table 2.15. Comparision of the accuracy of MEF (parameter set (w, m, p)=1400,5,9) with GRAIL2 predicting the band of width 40 *bp* (Data set: *DATA-WHOLE*).

Sequence	Length (*bp*)	MEF			GRAIL2			Winner[d]
		Sen.[a]	Spe.[b]	Sum[c]	Sen.[a]	Spe.[b]	Sum[c]	
1. Yeast								
SCCHRIII[e]	99750	89.5	81.2	170.7	50.9	79.2	130.1	MEF
YSCCHRVIN[e]	99750	78.0	78.4	156.4	48.0	66.1	114.1	MEF
SCCHRXVI[e]	99750	85.7	81.6	167.3	47.6	69.8	117.4	MEF
2. Chloroplast								
CHMPXX[e]	99750	22.6	82.9	105.5	20.4	77.4	97.8	MEF
CHNTXX[e]	99750	43.0	84.3	127.3	21.5	59.4	80.9	MEF
CHOSXX[e]	99750	31.1	83.3	114.4	18.9	67.5	86.4	MEF
3. C. elegans								
CELC50C3	44733	34.8	66.0	100.8	42.4	60.0	102.4	GRAIL2
CELF44E2	33651	33.3	77.5	110.8	14.3	45.9	60.2	MEF
CELTWIMUSC	54962	29.0	67.8	96.8	64.5	41.8	106.3	GRAIL2

(a) Sensitivity.
(b) Specificity.
(c) Sensitivity + specificity.
(d) More accurate algorithm in view of *Sum*(=sensitivity + specificity).
(e) A part (approximately 100,000 *bp*) of whole sequence is used because of the input limitation of GRAIL2.

Table 2.16. Comparision of the accuracy of MEF (parameter set (w, m, p)=1400,5,9) with GRAIL2 predicting the band of width 40 *bp* (Data set: *DATA-NEW*).

Sequence	Length (*bp*)	MEF			GRAIL2			Winner[d]
		Sen.[a]	Spe.[b]	Sum[c]	Sen.[a]	Spe.[b]	Sum[c]	
Yeast								
SC4357	75317	75.0	86.7	161.7	55.6	80.6	136.2	MEF
SC4987	21330	80.0	89.7	169.7	30.0	50.0	80.0	MEF
SC8339	27559	80.0	84.7	164.7	60.0	89.5	149.5	MEF
SC8520X	41200	90.0	83.4	173.4	60.0	82.6	142.6	MEF
SC9150	43504	75.0	83.5	158.5	58.3	94.4	152.8	MEF
SC9168	42793	79.2	74.5	153.7	50.0	61.5	111.5	MEF
SC9402	38990	100.0	75.4	175.4	31.2	50.0	81.2	MEF
SC9959	40397	82.6	83.7	166.3	56.5	95.2	151.8	MEF
SCD8035	69023	79.5	84.6	164.1	41.0	87.8	128.8	MEF
SCD9461	78500	76.5	81.3	157.8	51.0	84.4	135.4	MEF
SCD9717	72119	81.8	82.9	164.7	45.5	66.7	112.1	MEF
SCE6592	42576	73.9	75.9	149.8	47.8	75.0	122.8	MEF
SCE8199	36872	77.3	78.1	155.4	40.9	80.0	120.9	MEF
SCE9537	66030	86.1	79.9	166.0	58.3	81.6	139.9	MEF
SCE9781	68302	75.8	79.4	155.2	36.4	61.0	97.3	MEF
SCE9871	62643	71.9	81.9	153.8	53.1	73.0	126.1	MEF
SCU12980[e]	99750	89.4	78.0	167.4	42.6	61.2	103.8	MEF
YSCH8179	44100	74.1	75.3	149.4	59.3	92.3	151.6	GRAIL2
YSCH9196	55069	88.5	80.8	169.3	50.0	67.7	117.7	MEF
YSCLTESPO	29410	100.0	84.0	184.0	64.3	94.1	158.4	MEF
C. elegans								
CELC32E8	43726	19.7	55.4	75.1	18.2	39.0	57.2	MEF
CELF12B6	29268	35.0	30.4	65.4	37.5	50.0	87.5	GRAIL2
CELF53G12	41553	30.2	38.5	68.7	28.6	35.2	63.8	MEF
CELT12F5	32707	42.9	51.3	94.2	51.4	55.9	107.3	GRAIL2
Heli. pylori								
HPAE000539	10631	72.7	89.1	161.8	54.5	100.0	154.5	MEF
HPAE000557	16693	73.3	91.7	165.0	60.0	100.0	160.0	MEF
HPAE000608	11074	40.0	100.0	140.0	50.0	85.7	135.7	MEF
HPAE000651	10028	63.6	93.8	157.4	36.4	100.0	136.4	MEF

(a) Sensitivity.
(b) Specificity.
(c) Sensitivity + specificity.
(d) More accurate algorithm in view of *Sum*(=sensitivity + specificity).
(e) A part (approximately 100,000 *bp*) of whole sequence is used because of the input limitation of GRAIL2.

2.7 Summary and Discussion

In this chapter, the AIC and BIC procedures were compared, as a first step, to consider a Marov chain model for DNA sequences. It was observed, by simulations, that the AIC procedure is a more reliable order selection method than the BIC procedure for describing DNA sequences that are usually high order chain with four states.

As a main topic of this chapter, the relationship between an original Markov chain and its expanded chain was investigated, emphasising different starting points for expanding the process.

By applying the observed Markov chain relationship to DNA sequences, it was found that the non-stationary Markov chain property exists in exon DNA sequences, while intron sequences can be represented as stationary Markov chains.

The Markov chain properties in DNA sequences provided a basis to develop the MEF algorithm for detecting exon regions. Substantial improvement was observed when compared with the method based on correlation. Also, it was suggested that the widely used GRAIL2 program for combining multiple lines of evidence can be improved by using the simple Markov chain property.

The MEF algorithm has several advantages. Since the algorithm is primarily based on the global measure searched by content, it can be applied to the sequences across the phylogenetic spectrum and it is less affected by sequencing errors. Furthermore, the MEF algorithm is a fast and simple procedure. Its major limitation would be the unsatisfactory performance for relatively short exon sequence (especially, less than 200 *bp* in length) due to the use of a first-order transition matrix with 21 encoded amino acids. Another limitation includes the inability to precisely locate exon and intron boundaries.

Given these advantages and limitations, the suitable role of the MEF algorithm is to quickly scan the unknown DNA sequences and to indicate the potential exon sites. Therefore, in its present form, MEF can be used in concert with other algorithms that apply local property measurement such as the signal-based method.

CHAPTER 3
ZIPF'S LAW IN DNA SEQUENCES AND ITS RELATION TO MARKOV CHAIN

3.1 Background

3.1.1 Linguistic Features of DNA Sequences and Zipf's Law

A DNA sequence can be considered as a symbolic sequence of four-letter alphabets. The four letters are A, C, G, and T, indicating four nucleotides. So, a resonable analogy with natural human languages (English, German, French, etc.) comes to mind upon the first acquaintance with DNA texts. For example, it can be imagined that a DNA sequence consists of sentences (exon and intron regions), words (codons), punctuation marks (regulatory sites), and letters (nucleotides).

Natural language is a communication system characterized by the complex underlying structure (e.g., grammatical and semantic rules). This complexity draws a distinction between language and random text, where the random text means purely random symbolic sequences such as the materialwritten by a monkey at a typewriter. Properties reflecting the underlying structure of the communication system are called the *linguistic features* of the language

From the observation that a DNA sequence has a resemblance to language, the following question is naturally raised: Are linguistic features present in DNA sequences? In particular, is there a difference between exon and intron regions in DNA sequences

from linguistic viewpoint? In addition, what is the relation between linguistic features and Markov chain properties of DNA sequences? These questions motivated this chapter. If we can answer these questions, it would be another important clue to understand functions and structures of DNA sequences, especially, of the intron regions which are almost unknown.

One of the most famous quantative tools to measure the linguistic features is *Zipf's law* (Hubert, 1980). Zipf's law is an empirical rank-frequency relationship, which become first resonance in linguistics. Currently, two kinds of analyses are used to study Zipf's law: conventional and n-tuple analysis.

3.1.2 Conventional Zipf Analysis

In conventional Zipf analysis, the frequency of words presented in a given text is measured by counting the number of occurrences of each word throughout the text. The frequencies of all the words are then ordered from the most frequent to the least frequent. The position of each word in this ordered list is called its rank. Then the rank r of a word and corresponding frequency $w(r)$ (or relative frequency $p(r)$) satisfy the following empirical power-law (Zipf, 1949):

$$w(r) = \frac{C}{r^{\beta}}, \tag{3.1}$$

where C is an appropriate constant. In equation (3.1), $w(r)$ is replaced with $p(r)$ when the relative frequency is preferred. The exponent $\beta(> 0)$, which is called *Zipf's coefficient*, was found to be close to 1 in several texts written in various natural languages (Mandelbrot, 1983) by usual least square estimation with the $\log - \log$ transformation of word frequency versus its rank. Equation (3.1) is called *Zipf's law*. Figure 3.1 is an typical Zipf's plot, a plot of $\log(w(r))$ versus $\log(r)$, which usually shows reverse J-shape. Therefore, Zipf's law explains the linear region in the plot, or equivalently the upper tail part of the underlying distribution of word frequencies (if

present). Quite often, the Zipf's coefficient is estimated by fitting the linear region of the plot.

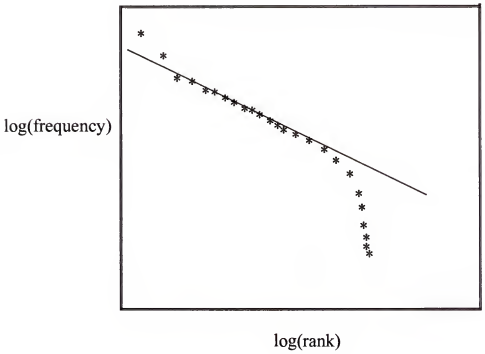

Figure 3.1. Example of typical Zipf's plot(*:observation, -:fitted line).

Since the publication of Zipf's work, there have been several attempts to prove his law. Here, to *prove* means to identify the population characteristics from which we can deduce Zipf's law. Previous studies will be discussed in Section 2.

In various areas, including linguistics, it is common that a plot of frequency against its rank yields a surprisingly close fit to Zipf's law. Here are a few examples: (1) a plot of the population of the r^{th} largest city (Berry and Garrison, 1958), (2) a plot of the income of the r^{th} richest family (Mandelbrot, 1960), (3) a plot of the r^{th} number of authors contributing a given number of papers to a journal or journals during a given period (Simon, 1955).

The current research focuses on Zipf's law in view of linguistics and its application to DNA sequences. However, the result can be generally applied to other areas as well.

3.1.3 n-tuple Zipf Analysis

Conventional Zipf analysis has a deficiency in the case of non-natural languages (e.g., technical language such as DNA sequence). Li (1992) showed that Zipf's law can emerge in a purely random symbolic sequence if one character is defined as a word delimiter and conventional Zipf analysis is applied. For example, if $\{AQEDACCDE \cdots\}$ is a random sequence and $\{E\}$ is defined as a word delimiter, then, by conventional Zipf analysis, $AQ, DACCD, \cdots$ are words. In this non-natural language, we may observe $\hat{\beta} > 0$ even if the sequence is purely random. Hence, in the analysis of non-natural languages, while the observation of the power-law behavior in a conventional Zipf analysis is necessary, it is not sufficient to prove the existence of linguistic features.

n-tuple Zipf analysis is used to overcome the deficiency of conventional analysis. The difference between conventional and n-tuple analysis is only in the counting method of occurrences of each word. In n-tuple Zipf analysis, the frequencies of substrings (words) of length n is measured by shifting progressively by 1 character a window of length n over the text (Mantegna et al., 1994), while the semantic units (real words) are counted in conventional Zipf analysis. Here, n is called *word length*.

The estimates of Zipf's coefficients from conventional and n-tuple analysis are usually different, although the similar reverse J-shapes to that shown in Figure 3.1 appear in Zipf's plot. Mantegna et al. (1995) applied the two Zipf's methods to English texts comprising approximately 10^6 words, which were selected from an encyclopedia. The estimate of Zipf's coefficient ($\hat{\beta}$) was 0.85 by conventional analysis of real words. On the other hand, they found $\hat{\beta}$=0.57 when n=6 by applying n-tuple

Zipf analysis to the same text, in which they used 32 character alphabets (consisting of the 26 letters of English and 6 punctuation symbols). The difference in resulting $\hat{\beta}$'s may arise due to the enlargement of the vocabulary in n-tuple analysis: Only real words constitute the vocabulary in the conventional analysis, while all possible n-tuples compose the vocabulary in the n-tuple case.

However, Zipf's law does not emerge (i.e., $\hat{\beta} \approx 0$) in n-tuple Zipf analysis of a purely random symbolic sequence (Mantegna et al., 1995). Therefore, n-tuple Zipf analysis is a more reasonable indicator of linguistic features. In addition, the n-tuple method is practically more useful in the case of a technical language such as DNA sequence, because the elementary semantic unit (word) is not immediately recognizable. In this chapter, n-tuple Zipf analysis is used.

3.1.4 Purpose of Chapter 3

The purpose of this chapter is to investigate the linguistic features of DNA sequences using Zipf's law. In particular, the following two questions will be considered: (1) Are the linguistic features in exon regions different from those in intron regions when they are measured by Zipf's law? (2) Can Markov chain properties fully explain the linguistic features of DNA sequences (if present)?

Possible ways to answer these questions follow: (1) Compare two Zipf's coefficients from exon and intron sequence. (2) Compare the Zipf's coefficient from a real DNA sequence with one from an artificial sequence generated by the Markov chain properties.

Appropriate statistical procedures to compare two Zipf's coefficients are necessary. No previous methods are available, however, to test the equivalence of two Zipf's coefficients, even though proving and applying Zipf's law have received a great deal of attention.

The main goals of this chapter are as follows: First, a model is built to appropriately describe the characteristics of word frequencies that lead to Zipf's law. Second, a test procedure is developed to compare two Zipf's coefficients under this model. Finally, answers to the above questions is provided by applying the developed test statistic to DNA sequences.

This chapter is divided into seven sections. Section 2 provides a literature review of related topics, including the studies that prove Zipf's law and apply the law to DNA sequences. Section 3 is devoted to deriving Zipf's law under an assumed model. The test procedure is constructed in Section 4. In Section 5, the assumed model and the built test procedure are verified by simulations. The application of the developed procedure to DNA sequences is the topic of Section 6. Finally, Section 7 provides summary and discussion.

3.2 Literature Review

3.2.1 Origin of Zipf's law

Nearly half a centry ago, Zipf (1949) observed that equation (3.1) is hold at least approximately in a wide variety of phenomena, including word fequencies. Equation (3.1) is called the *rank-frequency Zipf's law.* Zipf also found a size-frequency law such that the proportion of words with w frequency is approximately proportional to $w^{-\delta}$, for some $\delta > 0$. Throughout this chapter, we focus on the rank-frequency law, because most recent literature (e.g., Mantegna et al., 1995; Perlin, 1996; Troll and Graben, 1998) uses this form.

Since the publication of Zipf's novel work, many studies have followed to identify the origin of Zipf's law, i.e., to find a principle or underlying distribution form of word frequencies which can lead to Zipf's law.

Mandelbrot's (1953) explanation of Zipf's law is based on a "least cost vocabulary" principle. Let the probabilities associated with the r^{th} most common word in a person's vocabulary be $p(r)$ such that $p(r) \geq p(r+1)$ and $\sum_{r=1}^{\infty} p(r) = 1$. Again, let $h(r)$ be the amount of effort to extract the r^{th} most common word from memory. If Q is the amount of information per unit of effort, then

$$Q = -\frac{\sum_{r=1}^{\infty} p(r) \log p(r)}{\sum_{r=1}^{\infty} h(r) p(r)}.$$

By using the usual Lagrangian multiplier technique to maximize Q with the assumption

$$h(r) \propto \log r + a, \tag{3.2}$$

for some constant a, he obtained a more general form of Zipf's law:

$$p(r) = c(r+a)^{-\beta}, \tag{3.3}$$

for some constants a and c.

Good (1957) suggested a modification of Mandelbrot's assumption (3.2): For larger r, allow not only for the effort in recalling the r^{th} most common word, but also for the effort in initially learning it. With the following assumption which replaces (3.2):

$$h(r) \propto (1 + \epsilon p(r)^{-1}) \log r + a,$$

where ϵ is a very small positive constant, a variant of (3.3) is derived:

$$p(r) = c(r+a)^{-\beta(1+\epsilon p(r)^{-1})}.$$

Hill (1974) attempted to find a distribution form of word frequencies from which we can expect Zipf's law, while Mandelbrot and Good considered the underlying

mechanism in language. The model proposed by Hill is a Bose-Einstein form of the classical occupancy model.

Suppose there are N units (total number of words in a text) to be allocated in m nonempty cells (m different words) with w_i units in the ith cell (i.e., w_i is frequency of ith word) so that $w_i \geq 1$, $i = 1, \cdots, m$, and $\sum_{i=1}^{m} w_i = N$. Let $\underline{W}=(W_1, \cdots, W_m)$ be a random vector and $\underline{w}=(w_1, \cdots, w_m)$ be a vector of realization, where w_i denotes ith word frequency as above. Denote the order statistics in descending order by $W_{(1)} \geq W_{(2)} \geq, \cdots, \geq W_{(m)}$. Then, by Bose-Einstein allocation scheme, the probability function of $\underline{W}$ is given by the following:

$$P[\underline{\mathbf{W}} = (w_1, \cdots, w_m)|m, N] = \begin{pmatrix} N-1 \\ m-1 \end{pmatrix}^{-1}.$$

Here, assume that m is random, given N, with $F_N(x) = P[mN^{-1} \leq x|N] \xrightarrow{D} F(x)$ as $N \to \infty$, where F is an absolutely continuous distribution function. Hill showed that, if F is chosen such that $F'(x) \sim \beta x^{\beta-1}$ as $x \to 0$, where $\beta > 0$, then

$$E\Big[\frac{W(r)}{\log N}\Big] \propto r^{-\beta},$$

for sufficiently large N. This is called the *weak form of the rank-frequency Zipf's law.*

The Bose-Einstein model and its more general allocation scheme were also considered in several other studies to get size-frequency Zipf's law. Hill (1970) derived a weak form of the size-frequency law under the same model. Hill and Woodroofe (1975) showed that the Bose-Einstein model yields convergence in probability to the size-frequency Zipf's law, which is called the *stronger form of Zipf's law.* Chen (1980) suggested a Dirichlet-multinomial urn model, a more general form of the Bose-Einstein model, to get the size-frequency Zipf's law.

Rouault (1978) derived a more powerful form of the rank-frequency Zipf's law than the stronger form in the sense of Hill and Woodroofe (1975). The model is

based on Karlin's (1967) infinite urn scheme. Assume that word frequencies are distributed according to the multinomial distribution with total observations N and cell probabilities $p_i (i = 1, \cdots, m)$, where $\sum_{i=1}^{m} p_i = 1$. Let $p(1) \geq, \cdots, \geq p(m)$ be the ordered cell probabilities. Define

$$\alpha(x) = max\{i : p_i \geq \frac{1}{x}\}$$

and assume that $\alpha(x)$ obeys the relationship

$$\alpha(x) = x^{\delta} L(x),$$

for some δ, $0 < \delta < 1$, where $L(x)$ satisfies $L(cx)/L(x) \to 1$ as $x \to \infty$ for each c. Rouault showed that, for each r,

$$\frac{W(r)}{N} \xrightarrow{\text{a.s.}} cr^{-\beta} \text{ as } N \to \infty,$$

where $\beta = 1/\delta$ with $\beta > 1$ and $c = L(\frac{1}{p(r)})^{-1/\delta}$ with the same definition of $W(r)$ as before.

The result is surprisingly strong. However, it is not so meaningful since the primary interest of the rank-frequency Zipf's law is in the range of $0 < \beta \leq 1$.

In the middle of 1990s, the origin of the rank-frequency Zipf's law attracted attention again.

Perlin (1996) used a variant of the central limit theroem to show that the log-normal distribution as the underlying distribution of word frequencies can lead to Zipf's law. Let Y_n be the random variable to observe the relative frequency of a word which is composed of at most n letters. He represented Y_n as the product of independent and identically distributed random variables X_i's:

$$Y_n = X_1 X_2 \cdots X_{R_n},$$

where X_i's are the random variables representing letter probabilities, and R_n is also a random variable to denote word length, with $0 \leq R_n \leq n$. Then, he showed that Y_n is asymptotically log-normally distributed as $n \to \infty$.

Also, he provided an asymptotic formula for the least square estimator of Zipf's coefficient by log-log transformation of Zipf's law in the upper tail of the log-normal distribution. Then, he showed that the R^2 (coefficient of determination) associated with the regression approaches 1 as $n \to \infty$, i.e., Zipf's law can be expected. However, he did not verify the crucial assumption that a word is a sequence of letters appearing independently.

Troll and Graben (1998) provided another origin of Zipf's law. For large size of the vocabulary, they showed that the underlying distribution of word frequencies leads to a Pareto distribution by making the lower frequency bound go to zero. Also, it was proven that the rank-frequency Zipf's law can be directly derived from the Pareto distribution for word frequencies. This result reveals an origin of Zipf's law without strong assumptions.

3.2.2 Application of Zipf's Law to DNA Sequences

The application of Zipf's law to DNA sequences is in its beginning stage, even though the law has been widely used for several decades in various fields, including linguistics, economics, and so on.

One of the remarkable studies of Zipf's law in DNA sequences is probably Mantegna et al.'s (1994) work. They considered several kinds of DNA sequences to apply Zipf's law and measured the coefficients of Zipf's law to investigate differences between the linguistic features of exon and intron regions. They found that the estimates of Zipf's coefficient in intron regions (approximately $\hat{\beta} \approx 0.36$) were larger than those in exon regions ($\hat{\beta} \approx 0.21$) by using n-tuple rank-frequency Zipf's analysis with word length 4. They concluded that intron regions bear more resemblance to a natural

language ($\hat{\beta} \approx 0.57$ in the case of their encyclopedia data set) than the exon regions, because the $\hat{\beta} \approx 0.36$ in intron regions is closer to the $\hat{\beta} \approx 0.57$ in a natural language.

Mantegna et al. (1995) provided, following their earlier study, a relationship between the linguistic features of DNA sequences and Markov chain properties. They compared the Zipf's plot from natural DNA sequences with the corresponding Zipf's plots from the first-order Markovian approximation in exon regions and intron regions respectively. Here, the Markovian approximation means that the simulated sequence is based on the first-order transition matrix of the corresponding natural DNA sequence. It was visually observed that the deviation from the first-order Markovian approximation is maximal in intron regions. On the basis of this result, they concluded that the linguistic features of intron regions cannot be fully explained by first-order Markov chain properties.

A potential problem, however, in the works of Mantegna et al. (1994, 1995) is that they have not performed a statistical test to compare two estimates of Zipf's coefficients. They simply compared two observed values by visually inspecting discrepancies in the two plots, because of the lack of appropriate test procedure. This is one of the motivations of the current chapter.

Troll and Graben (1998) examined the empirical distribution functions of word frequencies in natural languages and DNA sequences. For word length $n = 5, 6, 7$, the word frequencies of English and German text show an empirical distribution function close to Pareto distribution in the upper tail. On the other hand, they showed that the log-normal distribution is a better approximation to the empirical distribution of the word frequencies in a DNA sequence (Yeast chromosome III, locus S288C).

In general, Mantegna et al.'s (1994, 1995) observation on DNA sequences is not quite conclusive because they consider only one kind of organism and so the trend in the empirical distribution function may be different if they inspect other organism. However, their observation will be revisited in Section 5 to check whether the new

procedure developed in this research can be still applied in the situation of a log-normal distribution of DNA word frequencies.

3.3 The model and Derivation of Zipf's Law

3.3.1 The Model

As reviewed in Section 2, there have been several attempts to identify the underlying distribution of word frequencies which can lead to Zipf's law. Most of them, however, derived Zipf's law under non-trivial assumptions. This might be one of the reasons why a test procedure for Zipf's coefficients is not available. In this section, a new model for the underlying distribution of word frequencies is considered and it will be shown that we can expect Zipf's law under this model. Also, a test statistic for Zipf's coefficients will be developed under the model in the next section.

Basically, the linguistic features measured by Zipf's law have global statistical properties in the sense that common characteristic exists over a vast volume of text. For example, it would be meaningless to apply Zipf's law to only several pages of an English book and German book if we want to investigate the difference between the linguistic features of English and German. Most pervious studies typically inspected the whole of a book, encyclopedia, or quite long DNA sequence. In this case, there are a great deal of different words (vocabulary) with a much greater number of total words. For example, the vocabulary size in James Joyce's Ulysses is approximately 29,900 (McNeil, 1973) in view of conventional Zipf's analysis. If n-tuple Zipf analysis is considered, the possible vocabulary size is C^n, where C is the number of letters used in a language and n is the word length. Table 3.1 shows the possible vocabulary sizes in the case of a DNA sequence (4^n).

Table 3.1. Possible vocabulary size in DNA sequence.

Word length	$n=5$	$n=6$	$n=7$	$n=8$
Vocabulary size	1024	4096	16384	65536

Therefore, it would not be totally unreasonable to consider a continuous type distribution, in the sense of approximation, for describing word frequencies rather than the multinomial distribution, which is natural in this situation. Most recently, Troll and Graben (1998) found that the Pareto distribution function is well fitted to the empirical distribution function of word frequencies in language texts. The degree of fitting, however, appears quite poor for the lower tail part of the distribution. Hence, a more generalized Pareto type is considered herein as the new model for the underlying distribution of word frequencies.

Let the vector of realization $(W_1, \cdots, W_m)$=$(w_1, \cdots, w_m)$ denote the word frequencies in a DNA sequence (more generally, language text) with length l, where m denotes total number of different words appearing in the sequence and $N = \sum_{i=1}^{m} w_i$. Assume that $W_1, W_2, \cdots, W_m$ are a sequence of positive independent random variables with a common distribution function F such that

$$1 - F(w) = w^{-\alpha} L(w), \quad \alpha > 0, \tag{3.4}$$

where $L(w)$ is a slowly varing function at infinity:

$$\frac{L(\lambda w)}{L(w)} \to 1 \quad when \quad w \to \infty, \quad \text{for any } \lambda > 0.$$

The present model is now often restated as the assumption of regular variation at infinity of $1 - F$ with index $-\alpha$ (Bingham et al., 1987). Denoting the order statistics

in descending order based on m observations $W_1, W_2, \cdots, W_m$ by

$$W_{(1)} \geq W_{(2)} \geq, \cdots, \geq W_{(r)} \geq, \cdots, \geq W_{(m)},$$

let

$$U(w) = inf\{y : F(y) \geq 1 - 1/w\}, \quad w > 1,$$

be the quantile function of F.

At the end of this section, it is shown that we can expect Zipf's law under model (3.4). Before proving the derivation of Zipf's law, we consider whether the above assumption is plausible for describing the distribution of word frequenices in DNA sequences.

First, the random variables W_i $(i = 1, \cdots, m)$ in the model are, strictly speaking, not independent. In particular, they obey the equality $\sum_{i=1}^{m} w_i = N$. However, it is well known (e.g., Günther et al., 1996) that the dependence vanishes for $N - m >> 1$ and $m >> 1$, which is the case of interest to us.

Second, the most important assumption is that $W_1, \cdots, W_m$ are from a common distribution function satisfying (3.4). To check this assumption, we can restate (3.4) as an equivalent model. More specifically, by Karamata's representation theroem (Bingham et al., 1987), model (3.4) is equivalent to the existence of functions c and b such that $c(w) \to c$ as $w \to \infty$ and $b(1/w) \to 0$ as $w \to \infty$, and that

$$U(w) = w^{\frac{1}{\alpha}} \bar{L}(w), \quad w > 1 \tag{3.5}$$

with

$$\bar{L}(w) = c(w) exp\Big(\int_{\frac{1}{w}}^{1} \frac{b(u)}{u} du \Big), \quad w > 1.$$

Now, by using Pareto quantile plot (Beirlant et al., 1996), this assumption can be checked. Before considering (3.4) or (3.5), we assume the Pareto distribution as the submodel of (3.4) (i.e., $F(w) = 1 - w^{-\alpha}$, $w > 1$). Then, the Pareto quantile plot can be used as a basis for testing the goodness-of-fit hypothesis of Pareto behavior. As

log-transformed Pareto distributed random variables are exponentially distributed with parameter $1/\alpha$, the theoretical quantiles of standard exponential distribution should be in linear relationship to the corresponding empirical quantiles of the log-transformed observations, if the model is correct. Thus, one can visually check the hypothesis of Pareto behavior from a sample $W_1, \cdots, W_m$ by inspecting the scatter-plot with coordinates

$$\Big(-\log \frac{r-1}{m}, \quad \log W_{(r)} \Big), \quad r = 2, \cdots, m. \tag{3.6}$$

Here the $\log W_{(r)}$ is used as the $(\frac{(m-r+1)}{m})^{th}$ empirical quantile for the $\log U(\frac{m}{(r-1)})$, which must stand in linear relationship to the corresponding theoretical quantile $-\log \frac{(r-1)}{m}$ of standard exponential distribution. If the Pareto quantile plot is linear, then the slope of a fitted line provides an estimate of $1/\alpha$. Next, when the general model (3.4) or (3.5) is considered, it follows from the fact that $\log \bar{L}(w)/\log w \to 0$ as $w \to \infty$ that

$$\log U(w) \sim -\frac{1}{\alpha} \log \frac{1}{w} \quad \text{as} \quad w \to \infty.$$

Noting that $\log W_{(r)}$ is an estimator of $\log U(\frac{m}{(r-1)})$, the Pareto quantile plot of (3.6) will eventually be linear for larger w values (i.e., smaller r values), or equivalently the scatterplot with

$$\Big(\log r, \quad \log W_{(r)} \Big) \tag{3.7}$$

must show a linear relationship for smaller r. The faster $\log \bar{L}(w)/\log w$ tends to zero as $w \to \infty$, the clearer this phenomenon appears. However, scatterplot with (3.7) is nothing but a Zipf's plot. In Figure 3.2, a Zipf's plot from the DNA sequence SCCHRIII shows this ultimate linear relationship for smaller ranks. The validity of model (3.4) will be checked once more by simulations in Section 5.

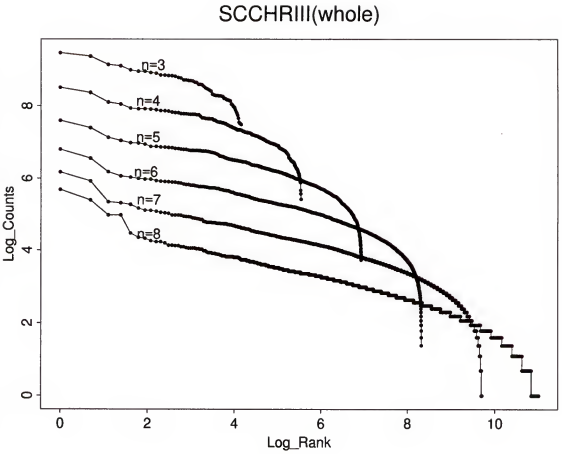

Figure 3.2. Zipf's plot for DNA sequence SCCHRIII.

3.3.2 Derivation of Zipf's Law

It is well known (e.g., Leadbetter et al., 1983) that the limiting distribution function of $W_{(r)}$ $(r = 1, \cdots, m)$ depends on the limiting behavior of the maximum of the sample$(W_{(1)})$. However, the nondegenerate limiting distribution function of $W_{(1)}$ has to be one of the Fréchet, Weibull, or Gumbel distributions, provided there exist constants $a_m \in R$ and $b_m > 0$ such that

$$F^m(a_m + b_m w) \to G(w), \quad m \to \infty \tag{3.8}$$

for every continuity point of the nondegenerate limiting distribution function G. However, Reiss (1989) shows that, if for every t,

$$sup\{w : F(w) < 1\} = \infty \tag{3.9}$$

and

$$\lim_{w \to \infty} \frac{1 - F(tw)}{1 - F(w)} = t^{-\alpha}, \quad \alpha > 0, \tag{3.10}$$

then G has to be Fréchet distribution:

$$G(w) = \begin{cases} exp(-w^{-\alpha}) & \text{if } w > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

In this case, one of the possible normalizing constants a_m and b_m can be chosen as

$$a_m = 0, \quad b_m = U(m). \tag{3.12}$$

Since model (3.4) satisfies (3.9) and (3.10), suitably normalized $W_{(1)}$ by (3.12) has the limiting distribution of (3.11). In this case, the limiting distribution of r^{th} order statistic, $W_{(r)}$, is given by (Reiss, 1989), for each r,

$$\lim_{m \to \infty} P[W_{(r)}/b_m] = \begin{cases} exp(-w^{-\alpha}) \sum_{t=0}^{r-1} \frac{(w^{-\alpha})^t}{t!}, & \text{if } w > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{3.13}$$

where $b_m = U(m)$.

Now, we consider the limiting behavior of $W_{(r)}$ in view of Zipf's law. The following theorem shows that we can expect Zipf's law in the region of the upper tail of model (3.4).

Theorem 3.1 Let $W_1, \cdots, W_m$ be a random sample from the distribution function F satisfying $1 - F(w) = w^{-\alpha}L(w)(\alpha > 0)$, where $L(w)$ is a slowly varing function, and let $W_{(1)} \geq, \cdots, \geq W_{(m)}$ be their order statistics. Then, for each r and sufficiently large m,

$$E[W_{(r)}/b_m] \propto r^{-\beta}, \tag{3.14}$$

where $\beta = 1/\alpha(\alpha, \beta > 0)$ and $b_m = inf\{y : F(y) \geq 1 - \frac{1}{m}\}$.

Remark We note that equation (3.14) holds in the region of the upper tail of the model due to the conditions of fixed r and large m.

Proof of theorem Using (3.13) and integration by part, for fixed r and $w > 0$, we obtain

$$\begin{aligned} \lim_{m\to\infty} P[W_{(r)}/b_m \leq w] &= exp(-w^{-\alpha}) \sum_{t=0}^{r-1} \frac{(w^{-\alpha})^t}{t!} \\ &= \int_{w^{-\alpha}}^{\infty} \frac{1}{(r-1)!} e^{-y} y^{r-1} dy \\ &= P[Y^{-\frac{1}{\alpha}} \leq w], \end{aligned}$$

where Y is a Gamma random variable with location parameter r and shape parameter 1. Note that $E[Y^{-\frac{1}{\alpha}}] \approx r^{-\frac{1}{\alpha}}$. The result follows.□

From Theroem 3.1, we expect that the relationship between $W_{(r)}$ and its fixed rank r should approximately obey Zipf's law with parameter β (> 0) for sufficiently large m. This is the weak form of Zipf's law in the sense of Hill (1974) and Chen (1980).

3.4 Asymptotic Normality of the Estimator of Zipf's Coefficient

The problem of estimating the characteristic parameter $\frac{1}{\alpha}(=\beta)$ has been treated by many authors (e.g., de Haan and Resnick, 1981; Bacro and Brito, 1993; Beirlant et al., 1996). However, most of them estimated the parameter under non-trivial assumptions on the slowly varing function L in (3.4). In the field of applying Zipf's law, a simple least square estimation by $\log - \log$ transformation is conducted to the linear part of Zipf's plot as mentioned in Section 1. In this section, therefore, the asymptotic normality of the least square estimator fitted to the upper tail region of distribution without any further assumption on the slowly varing function L, will be proven.

From the equation

$$E[\log W_{(r)}] = C - \beta \log r, \quad r = 1, \cdots, k,$$

where C is some constant and k is a fixed number ($\leq m$), the least square estimator of $-\beta$ is given by

$$\hat{\beta}_{m,k} = \frac{\sum_{i=1}^{k} (\log i - \overline{\log r}) \log W_{(i)}}{\sum_{i=1}^{k} (\log i - \overline{\log r})^2}, \tag{3.15}$$

where $\overline{\log r} = \frac{\sum_{i=1}^{k} \log i}{k}$. In this section, the asymptotic normality of $\hat{\beta}_{m,k}$ is considered when $W_{(1)}, \cdots, W_{(m)}$ are order statistics of the random samples $W_1, \cdots, W_m$ from (3.4).

On the asymptotic normality of linear combinations of the order statistics, several studies have been devoted to its development (e.g., Csörgö and Mason, 1985; Csörgö et al., 1988; Mason and Shorack, 1990; Csörgö et al., 1991; Mason and Shorack, 1992; Viharos, 1993). Recently, Viharos (1995) provided the following theorem.

Theorem 3.2 (Viharos, 1995) Let $W_{(1)} \geq W_{(2)} \geq, \cdots, \geq W_{(m)}$ be order statistics of m independent random variables with common distribution function F. Assume that $1 \leq k \leq m, \quad k \to \infty \quad \text{and} \quad k/m \to 0 \quad \text{as} \quad m \to \infty$. If the following conditions (A, B, and C) hold, then for all $\rho > -\frac{1}{2}$,

$$\frac{m^{\rho}((1+\rho)(1+2\rho)/(2\alpha^{-2}))^{1/2}}{l(k/m)k^{\rho+\frac{1}{2}}}\Big\{\sum_{i=1}^{k} d_{(i)} \log W_{(i)} - \mu_m(\bar{H})\Big\} \xrightarrow{D} N(0,1), \tag{3.16}$$

as $m \to \infty$, where

$$\begin{aligned} \mu_m(\bar{H}) &= -m\int_{\frac{1}{m}}^{\frac{k}{m}} J(1-u)\bar{H}(u)du - d_{(1)}\bar{H}(\frac{1}{m}), \qquad (3.17) \\ \bar{H}(s) &= -\log U(\frac{1}{s}), \quad s > 1, \end{aligned}$$

and ρ, α, k, and $J()$ are defined in the following conditions:

Condition A: There exist normalizing constants $a_m \in R$ and $b_m > 0$ such that

$$\lim_{m\to\infty} P[{b_m}^{-1}(W_{(1)} - a_m) \leq w] = K_\alpha(w)$$

for all w with $K_\alpha(w)$ necessary being an extreme value distribution function: $K_\alpha(w) = exp(-(1+\frac{1}{\alpha}w)^{-\alpha})$, where $\alpha \in R$ is a parameter, w is such that $1+\frac{1}{\alpha}w > 0$, and $(1+\frac{1}{\alpha})^{-\alpha}$ is interpreted as e^{-w} if $\alpha = 0$.

Condition B: The weights $d_{(i)}$ are of the form

$$d_{(i)} = m\int_{(i-1)/m}^{i/m} J(t)dt, \quad 1 \leq i \leq m,$$

for some continuous function J defined on $(0,1)$ which satisfies the following conditions:

(B-1) There exists a constant $0 < \mu < 1$ such that the function J is Lipschitz on $[1-\mu, 1-\delta]$ for all $0 < \delta < \mu$.

(B-2) There exists a constant $-\infty < \rho < \infty$ such that $J(1-t) = t^{\rho}l(t)$ on $(0,1)$ for some function $l()$ slowly varying at 0 and $l'(t) = t^{-1}l(t)\epsilon(t)$ on some $(0,\delta)$ with a continuous function $\epsilon()$ for which $\epsilon(t) \to 0$ as $t \to 0$.

(B-3) Suppose that $\rho > -1$ and for all $M > 1$,

$$\sup_{1/M<y<M} \left| \int_0^y \frac{(l(u/m) - l(y/m))u^{\rho}}{l(y/m)y^{\rho}} du \right| \to 0 \quad \text{as} \quad m \to \infty.$$

Condition C: $\lim_{s \downarrow 0} = d_0$, for some $0 < d_0 < \infty$

Proof see Viharos (1995). □

Now, by using Theroem 3.2, the asymptotic normality of the least square estimator of Zipf's coefficient given by (3.15) is followed when the model (3.4) is assumed.

Theorem 3.3 Let $W_{(1)} \geq W_{(2)} \geq, \cdots, \geq W_{(m)}$ be order statistics of positive random sample $W_1, W_2, \cdots, W_m$ from distribution function F such that $1-F(w) = w^{-\alpha}L(w)$, $\alpha > 0$, $w > 0$, where $L(w)$ is a slowly varying function at infinity. If $1 \leq k \leq m$, $k \to \infty$ and $k/m \to 0$ as $m \to \infty$, then the least square estimator of the negative value of Zipf's coefficient $(-\beta)$, $\hat{\beta}_{m,k} = \frac{\sum_{i=1}^{k}(\log i - \overline{\log r}) \log W_{(i)}}{\sum_{i=1}^{k}(\log i - \overline{\log r})^2}$, has a limiting distribution such that

$$T_{m,k}(\beta) = \frac{(1/(2\beta^2))^{1/2}}{l(k/m)k^{1/2}} \left\{ \hat{\beta}_{m,k} - \mu_m(\bar{H}) \right\} \xrightarrow{D} N(0,1) \tag{3.18}$$

as $m \to \infty$, where

$$\beta = \frac{1}{\alpha}, \quad \alpha, \beta > 0,$$

$$\overline{\log r} = \frac{\sum_{i=1}^{k} \log i}{k},$$

$$
\begin{aligned}
l(t) &= \frac{\log m(1-t) - \overline{\log r} + 1}{\sum_{i=1}^{k}(\log i - \overline{\log r})^2}, \quad 0 < t < 1, \\
\mu_m(\bar{H}) &= -m \int_{1/m}^{k/m} \frac{(\log m(1-u) - \overline{\log r} + 1)\bar{H}(u)}{\sum_{i}^{k}(\log i - \overline{\log r})^2} du, \\
\bar{H}(s) &= -\log U(\frac{1}{s}), \quad s > 1.
\end{aligned}
$$

Proof It is sufficient to check that the conditions in Theorem 3.2 are satisfied:

Condition A: From Section 3, the limiting distribution of appropriately normalized $W_{(1)}$ from the distribution function, F, given in the theroem has the form of

$$G(w) = exp(-w^{-\alpha}), \quad \alpha > 0, w > 0,$$

However, $G()$ is a special case of $K_\alpha(w)$ in Theroem 3.2 with $G(w) = K_\alpha(\alpha(w-1))$. Therefore, Condition A is satisfied.

Condition B: We want continuous function J on $(0,1)$ such that

$$\frac{\log i - \overline{\log r}}{\sum_{i=1}^{k}(\log i - \overline{\log r})^2} = m \int_{(i-1)/m}^{i/m} J(t)dt, \quad 1 \le i \le m.$$

Then, $J(t)$ can be approximated as

$$J(t) = \frac{\log mt - \overline{\log r} + 1}{\sum_{i=1}^{k}(\log i - \overline{\log r})^2}, \quad 0 < t < 1.$$

(B-1) Take μ such that $m(1-\mu) \ge 1$. Then, for all $0 < \delta < \mu$ and $x, y \in [1-\mu, 1-\delta]$,

$$
\begin{aligned}
|J(x) - J(y)| &= \frac{|\log mx - \log my|}{\sum_{i=1}^{k}(\log i - \overline{\log r})^2} \\
&\le \frac{m|x-y|}{\sum_{i=1}^{k}(\log i - \overline{\log r})^2}
\end{aligned}
$$

Thus, J is Lipschitz on $[1-\mu, 1-\delta]$.

(B-2) Take $\rho = 0$. Then, for $0 < t < 1$,

$$\begin{aligned} J(1-t) &= \frac{\log m(1-t) - \overline{\log r} + 1}{\sum_{i=1}^{k} (\log i - \overline{\log r})^2} \\ &:= l(t). \end{aligned}$$

For any $\lambda > 0$,

$$\frac{l(\lambda t)}{l(t)} = \frac{\log m(1-\lambda t) - \overline{\log r} + 1}{\log m(1-t) - \overline{\log r} + 1} \to 1 \quad as \quad t \to 0.$$

So, $l()$ is slowly varying at 0. Also, we can express

$$l'(t) = t^{-1}\Big[\frac{\log m(1-t) - \overline{\log r} + 1}{\sum_{i=1}^{k} (\log i - \overline{\log r})^2}\Big]\Big[\frac{-t}{1-t}(\log m(1-t) - \overline{\log r} + 1)^{-1}\Big].$$

Let $\epsilon(t) = -\frac{t}{1-t}(\log m(1-t) - \overline{\log r} + 1)^{-1}$. Then $\epsilon()$ is a continous function on $(0,1)$ with $\epsilon(t) \to 0$ as $t \to 0$.

(B-3) For all $M > 1$,

$$\begin{aligned} & \sup_{1/M<y<M} \Big| \int_0^y \frac{l(u/m) - l(y/m)}{l(y/m)} du \Big| \\ = & \sup_{1/M<y<M} \Big| \int_0^y \Big(\frac{\log m(1-u/m) - \overline{\log r} + 1}{\log m(1-y/m) - \overline{\log r} + 1} - 1 \Big) du \Big| \\ \to & \quad 0 \quad \text{as} \quad m \to \infty, \end{aligned}$$

since $\frac{\log m(1-M_2/m) - \overline{\log r} + 1}{\log m(1-M_1/m) - \overline{\log r} + 1} \to 1$ as $m \to \infty$ for every fixed $M_1 > 0$ and $M_2 > 0$.

Condition C:

$$\begin{aligned} \lim_{s \downarrow 0} J(1-s) &= \frac{\log m - \overline{\log r} + 1}{\sum_{i=1}^{k} (\log i - \overline{\log r})^2} \\ &:= d_0, \end{aligned}$$

where $0 < d_0 < \infty$. Hence, Condition C is satisfied.

Thus, the theorem is followed. □

From Theorem 3.3, under the hypothesis $H_0 : \beta_1 = \beta_2 (= \beta_0$, say), two estimates ($\hat{\beta}^{(1)}_{m,k}$ and $\hat{\beta}^{(2)}_{m,k}$) from two indepenent samples with equal m and k have

$$T_{m,k}(\beta_0) = \frac{\sum_{i=1}^{k}(\log i - \overline{\log r})^2 (1/(4\beta_0^2))^{1/2}}{(\log m(1-k/m) - \overline{\log r} + 1)k^{1/2}} \{\hat{\beta}^{(1)}_{m,k} - \hat{\beta}^{(2)}_{m,k}\} \tag{3.19}$$

$$\xrightarrow{D} N(0,1) \quad \text{as} \quad m \to \infty$$

However, $T_{m,k}(\beta_0)$ in (3.19) depends on the choice of k, which is called the *number of regressed observations*. One possible way to determine k is to restrict consideration to tests that control Type I error probability at a specified level α^*:

$$P[|T_{m,k}(\beta_0)| > Z_{1-\alpha^*/2}] = \alpha^*, \tag{3.20}$$

where $Z_{1-\alpha^*/2}$ is the $(\alpha^*/2)$ upper percentile of the standard normal distribution.

Determination of k for Zipf's Law

For application of Zipf's law to DNA sequences, k is determined by simulation for word length $n = 5, 6, 7, 8$, where k is the number of regressed observations included in the estimation of β.

For simulation, two independent samples are generated based on the Pareto distribution $F(w) = 1 - w^{-\alpha}$, which is a submodel of (3.4). This is because the estimation of β is excuted over the linear part of Zipf's plot while the Zipf's plot reveals linearity in the case of the Pareto distribution (as shown in Section 3). Also, there is prior information that Zipf's coefficient $\beta (= \frac{1}{\alpha})$ ranges roughly between 0.3 and 0.7 from previous studies (Mantegna et al., 1994, 1995) and a pilot study in this research.

The simulation was conducted as follows:

1. Choose word length n and Zipf's coefficient β.

2. Select k using vocabulary size m given by Table 3.1.

3. Generate two sets of m random numbers from Pareto distribution with given parameter $\frac{1}{\beta}$.

4. Calculate $\hat{\beta}_{m,k}^{(1)}$ and $\hat{\beta}_{m,k}^{(2)}$ according to (3.15).

5. Determine the absolute value of test statistic $|T_{m,k}(\beta)|$ according to (3.19).

6. It is a failure if $|T_{m,k}(\beta)| > Z_{1-\frac{\alpha^*}{2}}$ with given Type I error α^*.

7. Repeat step 3-6 s times.

8. Calculate failure rate of s times (empirical Type I error rate).

9. Repeat step 2-8 by changing k until the empirical Type I error rate satisfies the given α^*.

It was found, from the simulation, that the empirical Type I error rate shows a monotonically increasing trend as k gets larger, with a fluctuation which can be decreased as simulation time s becomes larger. Table 3.2 demonstrates the determined k by such a simulation as $\beta = 0.5, s = 100$, and $\alpha^* = 0.05$. For the cases of $\beta = 0.3$ and $\beta = 0.7$, results are quite similar, i.e., almost invariant to the choice of β. Also, we note that the determined k's for $n = 5, 6, 7, 8$ are roughly in the range of the linear part of Zipf's plot in Figure 3.2. In the following sections, Table 3.2 will be used for the application.

Table 3.2. Number of regressed observations[a] (k) to estimate Zipf's coefficients (Type I error=0.05).

Word length	$n=5$	$n=6$	$n=7$	$n=8$
k	765	3020	12200	47000

(a) Refer to Table 3.1 for the vocabulary sizes.

3.5 Validity of Test Procedure and Model

3.5.1 Validity of Test Procedure under Pareto Type Model

In this subsection, under the assumption that model (3.4) is correct, we check whether the test procedure suggested by (3.19) and (3.20) is appropriate to compare two estimates of Zipf's coefficients. This can be accomplished by inspecting the distribution behavior of test statistic $T_{m,k}(\beta)$ for all possible estimates of β. If the distribution is similar to the standard normal distribution for moderately large m, then we can admit the validity of the test procedure.

The simulation is excuted to investigate the distribution of $T_{m,k}(\beta)$ as follows:

1. Set word length n and Zipf's coefficient β.

2. Use vocabulary size m (Table 3.1) and determined the number of regressed observations k (Table 3.2).

3. Generate two sets of m random numbers from the Pareto distribution with given parameter $\frac{1}{\beta}$.

4. Calculate $\hat{\beta}^{(1)}_{m,k}$ and $\hat{\beta}^{(2)}_{m,k}$ according to (3.15).

5. Observe the value of $T_{m,k}(\beta)$ according to (3.19).

6. Repeat step 3-5 s times.

Table 3.3 shows sample moments of $T_{m,k}(\beta)$ for each $n = 5, 6, 7, 8$ and $\beta = 0.3, 0.5, 0.7$ with $s = 100$. If the test procedure is appropriate, then mean≈ 0, standard deviation≈ 1, skewness≈ 0, and kurtosis≈ 0 are expected. From Table 3.3, we observe that the distribution of $T_{m,k}(\beta)$ is not far from the standard normal distribution for each set of β's. In other words, the test procedure is appropriate under model (3.4).

Table 3.3. Sample moments of test statistic under the Pareto type model.

β	Word length	Mean	Standard deviation	Skewness	Kurtosis
$\beta = 0.3$	$n = 5$	0.1402	1.1454	-0.1957	0.6639
	$n = 6$	0.1580	1.0670	-0.6052	0.1620
	$n = 7$	-0.0149	0.9909	-0.1113	0.4540
	$n = 8$	0.1695	0.9470	-0.0612	-0.3614
$\beta = 0.5$	$n = 5$	-0.1455	1.1938	0.0370	0.5878
	$n = 6$	0.0494	0.9800	0.3179	-0.2376
	$n = 7$	0.0200	1.0736	0.2805	-0.2930
	$n = 8$	0.1695	0.9470	-0.0612	-0.3614
$\beta = 0.7$	$n = 5$	0.0346	0.9943	-0.0647	-0.4587
	$n = 6$	-0.0589	1.0765	0.2651	-0.6055
	$n = 7$	0.1098	1.0077	-0.1941	-0.3764
	$n = 8$	-0.0882	0.8700	-0.6441	0.3009

3.5.2 Validity of Test Procedure under log-normal Model

Troll and Graben (1998) suggested, by inspecting one DNA sequence (S288C), that the empirical distribution function of word frequencies in DNA sequences is well fitted to the log-normal distribution, even though natural languages follow the Pareto distribution. This subsection is devoted to the possibility that the underlying distribution of word frequencies is log-normal rather than the Pareto type distribution.

First, we check whether Zipf's law can be applied to the log-normal situation. Let the vector of realization $(W_1, \cdots, W_m)$=$(w_1, \cdots, w_m)$ denote word frequencies in a DNA sequence. Assume that $W_1, \cdots, W_m$ are a sequence of independent random variables with a common log-normal distribution such that

$$E[\log W_i] = 0, \quad Var[\log W_i] = 1, \quad i = 1, \cdots, m.$$

Also, let $W_{(1)} \geq W_{(2)} \geq, \cdots, \geq W_{(r)} \geq, \cdots, \geq W_{(m)}$ be their order statistics. Then, due to Leadbetter et al. (1983), for fixed r,

$$\lim_{m\to\infty} P\Big[\frac{\log W_{(r)} - a_m}{b_m} \leq w\Big] = e^{-e^{-x}} \sum_{t=0}^{r-1} \frac{(e^{-x})^t}{t!} \tag{3.21}$$

with

$$\begin{aligned} a_m &= (2\log m)^{1/2} - \frac{1}{2}(2\log m)^{-1/2}(\log\log m + \log 4\pi), \\ b_m &= (2\log m)^{-1/2}. \end{aligned}$$

Then, by using the technique similar to the proof of Theorem 3.1,

$$\frac{\log W_{(r)} - a_m}{b_m} \xrightarrow{D} -\log Y, \quad \text{as} \quad m \to \infty, \tag{3.22}$$

where Y is a Gamma random variable with location parameter r and shape parameter 1. However, $E[-\log Y] \approx -\log r$. Thus, $E[\frac{\log W_{(r)} - a_m}{b_m}] \propto -\log r$ for large m, i.e., Zipf's law with coefficient 1.

More generally, if the W_i's $(i = 1, 2, ..., m)$ are independent random variables from the log-normal distribution such that $E[\log W_i] = \mu$ and $Var[\log W_i] = \sigma^2$, then for large m and fixed r (i.e., for upper tail part of the model),

$$E[\frac{\log W_{(r)} - a'_m}{b'_m}] \propto -\sigma \log r \tag{3.23}$$

with appropriate normalizing constants a'_m and b'_m.

Thus, Zipf's law is expected for the upper tail of the log-normal distribution for sufficiently large m. This is also the weak form of Zipf's law in the sense of Hill (1974) and Chen (1980).

Next, we check the validity of the test procedure (3.19) in the log-normal situation, even though the test statistic is designed for Pareto type model (3.4). Simulation is again conducted to investigate the distribution of $T_{m,k}(\beta)$.

The simulation procedure is identical to the one in section 3.5.1 except that "the Pareto distribution with parameter $\frac{1}{\beta}$" is replaced with "the log-normal distribution with $E[\log W_i] = \mu$ and $Var[\log W_i] = \sigma^{2}$". Here, the average of two estimates $\hat{\beta}^{(1)}_{m,k}$ and $\hat{\beta}^{(2)}_{m,k}$ is used for β in $T_{m,k}(\beta)$. We note, from (3.23), that σ plays a role in the Zipf's coefficient while the location parameter μ has little effect on the behavior of $T_{m,k}(\beta)$. Table 3.4 shows the sample moments of $T_{m,k}(\beta)$ when $\mu = 0.5$ and $\sigma = 0.5$.

All sample moments, except sample standard deviations, are quite close to 0 for each n. Also, the sample standard deviations are not far from 1 for large n, even though they are less than 1. Other choices of $\sigma = 0.3$ and $\sigma = 0.7$ show similar results. Figure 3.3 shows the differences of the cumulative probabilities between the standard log-normal distribution and various Pareto distributions. The similar behavior in upper tail parts of both distributions was observed. This might be one of factors to force the sample moments of the test statistic under the log-normal assumption to be distributed as standard normal variate, because the test procedure uses only the upper tail part of the observations. These results suggest that the test

statistic can be applied with moderate degree of error to the situation in which the true underlying distribution of word frequencies is log-normal.

Table 3.4. Sample moments of the test statistic under the log-normal model.

Word length	Mean	Standard deviation	Skewness	Kurtosis
$n = 5$	0.0800	0.6052	-0.1465	0.3289
$n = 6$	0.0062	0.7073	-0.1766	-0.0161
$n = 7$	0.0143	0.8412	-0.1169	0.3664
$n = 8$	-0.0732	0.8603	0.3278	-0.2207

3.5.3 Validity of the Model: Case of DNA Sequences

In this subsection, the validity of model (3.4) is checked by using actual DNA sequences.

In Chapter 2, it was shown that DNA sequences can be well represented by a third-order Markov chain. That is, the simulated sequence based on the third-order transition matrix of a DNA sequence is not much different from the original DNA sequence. Therefore, the sample moments of $T_{m,k}(\beta)$, when it is applied to numerous sequences simulated in such a way, should show the shape of the standard normal distribution, provided the model (3.4) is correctly assumed for DNA sequences. Also, Subsection 3.5.1 has already provided the validity of $T_{m,k}(\beta)$ under the assumption that model (3.4) is true. This observation enables us to investigate the validity of

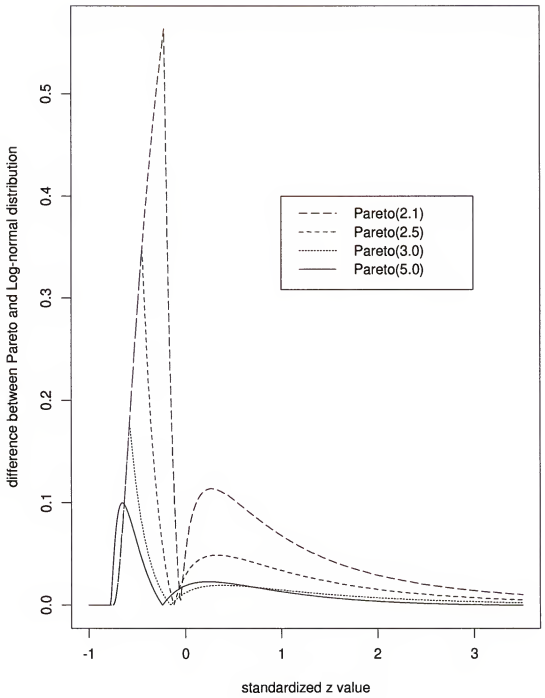

Figure 3.3. Difference in cumulative probability between Standard log-normal distribution and Pareto(β) distribution functions(β=0.21, β=2.5, β=3.0, β=5.0).

model (3.4), which was the assumed underlying distribution of word frequencies in DNA sequences.

The following simulation procedure is excuted to check validity:

1. Choose a DNA sequence.
2. Fix word length n and sequence length l.
3. Use the determined number of regressed observations k (Table 3.2).
4. Generate two sequences based on a third-order transition matrix of the DNA sequence.
5. Calculate $\hat{\beta}_{m,k}^{(1)}$ and $\hat{\beta}_{m,k}^{(2)}$ according to (3.15).
6. Determine $T_{m,k}(\beta)$ according to (3.19). Here, average of $\hat{\beta}_{m,k}^{(1)}$ and $\hat{\beta}_{m,k}^{(2)}$ is used for β in $T_{m,k}(\beta)$.
7. Repeat step 4-6 s times.

Table 3.5 shows sample moments of the test statistic $T_{m,k}(\beta)$ when a long DNA sequence SCCHRIII is used with $s = 100$. It is evident that the distribution of $T_{m,k}(\beta)$ behaves like the standard normal distribution. This implies that the model (3.4) is appropriate to describe word frequencies in DNA sequence.

3.6 Application of the Test Procedure to DNA Sequences

3.6.1 The Data

Three data sets, YEAST, CHLOROPLAST, and C.ELEGANS are used for the application. Each data set is composed of two sequences, exon and intron, of equal

Table 3.5. Sample moments of the test statistic for simulated DNA sequences based on SCCHRIII.

Word length	Seq. length(bp)[a]	Mean	Standard dev.[b]	Skewness	Kurtosis
$n = 5$	3080	0.0711	0.7021	0.5894	0.5084
$n = 6$	12390	0.0460	0.8988	-0.0117	0.2796
$n = 7$	49910	-0.0358	0.9279	0.0455	0.0678
$n = 8$	200060	0.0537	0.9794	0.0178	-0.3390

(a) Sequence length.
(b) Standard deviation.

length. The exon and intron sequences for each data set were taken from the following DNA sequences (see Table 2.6): (1) YEAST – SCCHRIII, YSCCHRVIN, SCCHRXVI, (2) CHLOROPLAST – CHMPXX, CHNTXX, CHOSXX, (3) C.ELEGANS – CELC50C3, CELF44E2, CELTWIMUSC.

From each of three DNA sequences in a data set, the exon regions were extracted and combined into one exon sequence, while the remaining fragments composed the intron sequence. In particular, exon and intron sequences in a data set were controlled to be equal length, because different length might cause a bias when we compare an exon sequence with an intron sequence. The length of the exon and intron sequences in the data sets is as follows: YEAST (210,210 bp), CHLOROPLAST (170,310 bp), and C.ELEGANS (50,820 bp).

3.6.2 Test for Differences between Exon and Intron Sequences

By applying the test procedure (3.19) to three data sets, we try to see the difference between exon and intron regions measured by Zipf's law. Here, word lengths 6 and 7

are considered since length 6 means 2 codons, which is attracting much attention in biological literature. Also, word length 7 is taken to be compared with word length 6. In addition, Table 3.2 is used as the number of regressed observations. Table 3.6 shows the estimates of Zipf's coefficient in exon and intron sequences. It can be concluded, based on the given data set, that in the cases of word length 6 and 7, the linguistic features of exon regions are different from those of intron regions.

Table 3.6. Comparison of Zipf's coefficients between exon and intron regions.

Word length	Data set	Estimated Zipf's coeff. exon[a]	intron[b]	P-value
n=6	YEAST	0.4254	0.5798	0.0000
	CHLOROPLAST	0.5716	0.6272	0.0092
	C.ELEGANS	0.5810	0.7019	0.0000
n=7	YEAST	0.4880	0.6404	0.0000
	CHLOROPLAST	0.6438	0.6854	0.0002
	C.ELEGANS	0.7219	0.8048	0.0000

(a) Estimated Zipf's coefficient from exon sequence.
(b) Estimated Zipf's coefficient from intron sequence.

3.6.3 Relationship between Linguistic Features and Markov Chain Properties

This subsection is devoted to answering the question: Can the Markov chain properties fully explain the linguistic features of DNA sequences? To answer this question, the simulation technique is used again. For each sequence in the data sets, simulated sequences are repeatedly generated based on the transition matrix

(independent, first, second, and third order are considered) of the original sequence. For each simulated sequence, the estimate of Zipf's coefficient is calculated. Then, we compare the estimate of Zipf's coefficient from the original sequence with the average of the estimates from the multiple simulated sequences (here, the simulation is 100 times). Equivalence of the two estimates implies that the Markov chain properties can explain linguistic features. This is because the simulated sequences are completely controlled by the transition matrix with a given Markov chain order.

Tables 3.7 - 3.10 demonstrate test results from the test procedure in (3.19) for independent, first-, second-, and third- order simulation respectively. From the tables, we observe that the first-order Markov chain can explain the linguistic features of intron regions, while a third-order representation is needed to explain the linguistic features of exon regions.

Table 3.7. Comparison of Zipf's coefficients between original DNA sequence and simulated sequence (Simulation order: independent).

Length	Data set	Exon			Intron		
		Zipf's coeff.			Zipf's coeff.		
		Orig.[a]	Simu.[b]	P-value	Orig.[a]	Simu.[b]	P-value
$n=6$	YEAST	0.4254	0.3351	0.0000	0.5798	0.5325	0.0170
	CHLOROPLAST	0.5716	0.4826	0.0000	0.6272	0.5392	0.0000
	C.ELEGANS	0.5810	0.3146	0.0000	0.7019	0.6080	0.0001
$n=7$	YEAST	0.4880	0.3841	0.0000	0.6404	0.5973	0.0000
	CHLOROPLAST	0.6438	0.5441	0.0000	0.6845	0.6042	0.0000
	C.ELEGANS	0.7219	0.4209	0.0000	0.8048	0.7264	0.0000

(a)Estimated Zipf's coefficient from original DNA sequence.
(b)Average of estimated Zipf's coefficients from 100 simulated DNA sequences.

Table 3.8. Comparison of Zipf's coefficients between original DNA sequence and simulated sequence (Simulation order: first).

Length	Data set	Exon			Intron		
		Zipf's coeff.			Zipf's coeff.		
		Orig.[a]	Simu.[b]	P-value	Orig.[a]	Simu.[b]	P-value
$n = 6$	YEAST	0.4254	0.3941	0.0324	0.5798	0.5582	0.2861
	CHLOROPLAST	0.5716	0.5475	0.2262	0.6272	0.6163	0.6239
	C.ELEGANS	0.5810	0.4959	0.0000	0.7019	0.6878	0.5670
$n = 7$	YEAST	0.4880	0.4475	0.0000	0.6404	0.6164	0.0238
	CHLOROPLAST	0.6438	0.6072	0.0005	0.6845	0.6746	0.3893
	C.ELEGANS	0.7219	0.6256	0.0000	0.8048	0.7887	0.2306

(a)Estimated Zipf's coefficient from original DNA sequence.
(b)Average of estimated Zipf's coefficients from 100 simulated DNA sequences.

Table 3.9. Comparison of Zipf's coefficients between original DNA sequence and simulated sequence (Simulation order: second).

Length	Data set	Exon			Intron		
		Zipf's coeff.			Zipf's coeff.		
		Orig.[a]	Simu.[b]	P-value	Orig.[a]	Simu.[b]	P-value
$n = 6$	YEAST	0.4254	0.4058	0.1858	0.5798	0.5754	0.8285
	CHLOROPLAST	0.5716	0.5552	0.4130	0.6272	0.6208	0.7764
	C.ELEGANS	0.5810	0.5422	0.0524	0.7019	0.6853	0.5004
$n = 7$	YEAST	0.4880	0.4623	0.0013	0.6404	0.6328	0.4810
	CHLOROPLAST	0.6438	0.6109	0.0202	0.6845	0.6787	0.6144
	C.ELEGANS	0.7219	0.6787	0.0025	0.8048	0.7900	0.2711

(a)Estimated Zipf's coefficient from original DNA sequence.
(b)Average of estimated Zipf's coefficients from 100 simulated DNA sequences.

Table 3.10. Comparison of Zipf's coefficients between original DNA sequence and simulated sequence (Simulation order: third).

Length	Data set	Exon			Intron		
		Zipf's coeff.			Zipf's coeff.		
		Orig.[a]	Simu.[b]	P-value	Orig.[a]	Simu.[b]	P-value
$n = 6$	YEAST	0.4254	0.4233	0.8938	0.5798	0.5789	0.9652
	CHLOROPLAST	0.5716	0.5677	0.8476	0.6272	0.6212	0.7903
	C.ELEGANS	0.5810	0.5629	0.3695	0.7019	0.6963	0.8200
$n = 7$	YEAST	0.4880	0.4808	0.3772	0.6404	0.6349	0.6075
	CHLOROPLAST	0.6438	0.6318	0.2643	0.6845	0.6771	0.5219
	C.ELEGANS	0.7219	0.7079	0.2465	0.8048	0.7969	0.5600

(a)Estimated Zipf's coefficient from original DNA sequence.
(b)Average of estimated Zipf's coefficients from 100 simulated DNA sequences.

3.7 Summary and Discussion

In this chapter, it was shown that Zipf's law can be lead when the Pareto type regular variation model is served as a underlying distribution of word frequencies. Also, a statistical tool to test the equivalence of two Zipf's coefficients was suggested. The developed procedure can be widely used in many areas, in which Zipf's law is applied, as well as linguistics.

By applying the test procedure to DNA sequences, it was found that the linguistic features of exon regions are different from those of intron regions. In particular, estimates of Zipf's coefficients from intron sequences are significantly greater than the ones from exon sequences in the cases of word length 6 and 7. Such a finding

agrees with Mantegna et al. (1994, 1995), even though they visually inspect the difference.

This chapter also revealed, through simulations, the relationship between linguistic features and Markovian properties of DNA sequences. In particular, the linguistic features of intron regions can be explained by a lower order Markov chain than those of exon regions. It may be interesting to compare with Mantegna et al.'s (1995) work, in which the deviation from the first-order Markovian approximation is maximal in intron regions in view of Zipf's behavior. However, they did not consider Zipf's coefficient but looked at the rank-frequency Zipf's graph itself. Therefore, it is inconclusive whether their work contradicts the result of this chapter.

In the previous chapter, it was already observed that intron sequences can be represented as stationary Markov chains while exon sequences can not be. This may imply that intron regions are closer to usual Markov chain and so linguistic features of intron regions could be explained by a lower order Marov chain than of exon region.

CHAPTER 4
CONCLUSION

The application of statistical techniques to DNA sequences is in its beginning stages. If the evolution or function of genes, proteins etc. obeyed only certain deterministic rules, then experimental biologists would be able to tell us exactly what these rules are. However, there is a stochastic or hidden structure in DNA sequences. The benefits of finding the hidden structure would be enormous as it may eventually assist experimental biologists in designing experiments and interpretating results. In addition, the search for determinism in DNA sequences may not evolve unless the hidden principles are discovered. This research is a step toward that direction.

In this dissertation, the relationship between the Markov chain order of DNA sequences and amino acid sequences were identified by means of the expanded Markov chain technique. In addition, it was found that the non-stationary Markov chain property exists in exon regions of a DNA sequence. While intron sequences can be represented as stationary Markov chain, exon sequences cannot be.

A new algorithm (MEF) to detect exon regions was developed based on the Markov chain property. The MEF algorithm is able to detect relatively long (approximately, longer than 800 *bp*) exon regions with a high degree of reliability. Since the Markov chain property is basically a global measurement, it can be applied to the sequences across phylogenetic spectrum and it is less affected by sequencing error. Furthermore, the MEF algorithm is a fast and simple procedure. Its major limitations would be

the unsatisfactory performance for relatively short exon regions and the inability to precisely locate exon and intron boundaries.

Considering these advantages and limitations, the MEF algorithm based on the simple Markov chain property seems to be a promising approach for quickly scanning unknown DNA sequences and indicating the potential exon regions. Thus, by combining MEF with other tools, such as signal-based measure, to find exon boundaries the precise location of exon segments might be possible. This is a natural next step of research.

With respect to the linguistic features of DNA sequences, the test procedure to compare two estimates of Zipf's coefficient were developed, assuming a Pareto type regular variation model for the underlying distribution of word frequencies. The validity of the model and the test procedure were verified by suitable methods. The procedure can be applied to various field, in which Zipf's law is applied, as well as linguistics.

It was also found that, in the given data sets, the linguistic features of intron regions are significantly different from those of exon regions in view of Zipf's law. In addition, the application of the test procedure to the simulated DNA sequences revealed, based on the experiments, that the first-order Markov chain properties can explain the linguistic features of intron regions, while third order is needed for exon regions. This observation may be another important clue for interpreting the function of intron regions, which is almost unknown.

Considering the stationary property of the Markov chain for intron regions, we could make a clearer explanation on the discrepancy of linguistic features between exon and intron regions, provided further research is followed to investigate the effect of the Markov chain on the underlying distribution of word frequencies.

REFERENCES

Anderson, T.W., and Goodman, L.A. 1957. Statistical Inference about Markov Chains. *Ann. Math. Statist.* 28:89–110.

Almagor, H. 1983. A Markov Analysis of DNA Sequences. *J. Theor. Biol.* 104:633–645.

Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control.* AC-19:716–723.

Arnold, J., Cuticchia, A.J., Newsome, D.A., Jennings, W.W., and Ivarie, R. 1987. Mono-Through Hexanucleotide Composition of the Sense Strand of Yeast DNA: A Markov Chain Analysis. *Nucleic Acids Research.* 15:2627–2638.

Arques, D.G., and Michel, C.J. 1987. Periodicities in Introns, *Nucleic Acids Research.* 15:7581–7592.

Bacro, J.N., and Brito, M. 1993. Strong Limiting Behaviour of a Simple Pareto-Index Estimator. *Statistics and Decisions.* 3:133–134.

Beirlant, J., Vynckier, P., and Teugels, J. 1996. Tail Index Estimation, Pareto Quantile Plots, and Regression Diagnostics. *Journal of the American Statistical Association.* 91:1659–1667.

Berry, B., and Garrison, W. 1958. Alternate Explanations of Urban Rank Size Relations. *Annals of the Association of American Geographers.* 48:83–91.

Bingham, N., Goldie, C., and Teugels, J. 1987. Regular Variation. Cambridge: Cambridge University Press.

Bishop, D.T., Williamson, J.A. and Skolnik, M.H. 1983. A Model for Restriction Fragment Length Distributions. *Amer. J. Hum. Genet.* 35:795–815.

Blaisdell, B.E. 1985. Markov Chain Analysis Finds a Significant Influence of Neighboring Bases on the Occurrence of a Base in Eucaryotic Nuclear DNA Sequences Both Protein–Coding and Noncoding. *J. Mol. Evol.* 21:278–288.

Chen, W.-C. 1980. On the Weak Form of Zipf's Law. *J. Appl. Prob.* 17:611–622.

Chou, P.Y., and Fasman, G.D. 1987. Empirical Predictions of Protein Conformation. *Ann. Rev. Biochem.* 47:251–276.

Csőrgő, S., and Mason, D.M. 1985. Central Limit Theorems for Sums of Extreme Values. *Math. Proc. Cambridge Philos. Soc.* 98:547–558.

Csőrgő, S., and Mason, D.M. 1988. The Asymptotic Distribution of Trimmed Sums. *Ann. Probab.* 16:672–699.

Csőrgő, S., Haeusler, E., and Mason, D.M. 1991. The Asymptotic Distribution of Extreme Sums. *Ann. Probab.* 19:783–811.

Cuticchia, A.J., Ivarie, R., and Arnold J. 1992. The Application of Markov Chain Analysis to Oligonucleotide Prediction and Physical Mapping to Drosophila Melanogaster. *Nucleic Acids Research.* 20:3651–3657.

de Haan, L., and Resnick, S. 1980. A Simple Asymptotic Estimate for the Index of a Stable Distribution. *Journal of the Royal Statistical Society, Ser. B.* 42:83–87

Fickett, T.W., and Tung, C. (1992). Assessment of Protein Coding Measures. *Nucleic Acids Research.* 20:6441–6450.

Findley, D.F. 1992. Counter Examples to Parsimony and BIC. *Ann. Inst. Statist. Math.* 43:505–514.

Gates, P., and Tong, H. 1976. On Markov Chain Modeling to Some Weather Data. *Journal of Applied Meteorology.* 15:1145–1151.

Good, I.J. 1957. Distribution of Word Frequencies. *Nature.* 179:595.

Gűnther, R., Levitin, L., Schapiro, B., and Wagner, P. 1996. Zipf's Law and the Effect of Ranking on Probability Distributions. *International Journal of Theoretical Physics.* 35:395–417.

Gut, A. 1987. Stopped Random Walks: Limit Theorems and Applications. New York: Springer-Verlag.

Hill, B.M. 1970. Zipf's Law and Prior Distributions for the Composition of a Population. *Journal of the American Statistical Association,* 65:1220–1232.

Hill, B.M. 1974. The Rank-Frequency Form of Zipf's Law. *Journal of the American Statistical Association.* 69:1017–1026.

Hill, B.M., and Woodroofe, M. 1975. Stranger Form of Zipf's Law. *Journal of the American Statistical Association.* 70:212–219.

Hubert, J.J. 1980. Linguistic Indicators. *Social Indicators Research.* 8:223–255.

Karlin, S. 1967. Central Limit Theorems for Certain Infinite Urn Schemes. *J. of Math. and Mech.* 17:373–401.

Karlin, S., and Brendel, V. 1993. Patchiness and Correlations in DNA Sequences. *Science.* 259:677–680.

Katz, R.W. 1981. On Some Criteria for Estimating the Order of a Markov Chain. *Technometrics.* 23:243–249.

Kelly, F.P. 1982. Markov Functions of a Markov Chain. *Sankhya: The Indian Journal of Statistics, Series A.* 44:372–379.

Kemeny, J.G., and Snell, J. L. 1976. Finite Markov Chains. New York: Springer-Verlag.

Kullback, S., Kupperman, M., and Ku, H.H. 1962. Test for Contingency Tables and Markov Chains. *Technometrics.* 4:573–608.

Lange, K. 1997. Mathematical and Statistical Methods for Genetic Analysis. New York: Springer–Verlag.

Leadbetter, M., Lindgren, G., and Rootzen, H. 1983. Extremes and Related Properties of Random Sequences and Processes, New York: Springer-Verlag.

Li, W. 1992. Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. *IEEE transaction on Information Theory* 38:1842–1845.

Lindqvist, B. 1978. On the Loss of Information Incurred by Lumping States of a Markov Chain. *Scand. J. Statist.* 5:92–98.

Mandelbrot, B.B. 1953. An Informational Theory of the Statistical Structure of Language. In: Communication Theory, W. Jackson (ed.). London: Butterworths.

Mandelbrot, B.B. 1960. The Pareto-Levy Law and the Distribution of Income. *International Economic Review.* 1:79–106.

Mandelbrot, B.B. 1983. The Fractal Geometry of Nature. New York: Freeman.

Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simon, M., and Stanley, H.E. 1994. Linguistic Features of Non–Coding DNA Sequences. *Physical Review Letters.* 73:3169–3172.

Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simon, M., and Stanley, H.E. 1995. Systemic Analysis of Coding and Noncoding DNA Sequences Using Methods of Statistical Linguistics. *Physical Review E.* 52:2939–2950.

Mason, D.M., and Shorack, G.R. 1990. Necessary and Sufficient Conditions for Asymptotic Normality of Trimmed L-statistics. *J. Statist. Plann. Inference.* 25:111–139.

Mason, D.M., and Shorack, G.R. 1992. Necessary and Sufficient Conditions for Asymptotic Normality of L-Statistics. *Ann. Probab.* 20:1779–1804.

McNeil, D.R. 1973. Estimating an Author's Vocabulary. *Journal of the American Statistical Association.* 68:92-96.

Norberg, T. 1997. On the Time a Markov Chain Spends in a Lumped State. *J. Appl. Prob.* 34:340-345.

Ossadnik, S.M., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Peng, C.-K., Simons, M., and Stanley, H.E. 1994. Correlation Approach to Identify Coding Regions in DNA Sequences. *Biophyical Journal.* 67:64–70.

Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Halvin, S., Sciortino, F., Simons, M., and Stanley, H.E. 1992. Long-Range Correlations in Nucleotide Sequences. *Nature.* 356:168–170.

Peng, C.-K., Buldyrev, S.V., Halvin, S., Simons, M., Stanley, H.E., and Goldberger, A.L. 1994. On the Mosaic Organization of DNA Sequences. *Physical Review E.* 49:1685–1689.

Perline, R. 1996. Zipf's Law, the Central Limit Theorem, and the Random Division of the Unit Interval. *Physical Review E.* 54:220–223.

Phillips, G.J., Arnold J., and Ivarie, R. 1987. Mono-Through Hexanucleotide Composition of the Escherichia Coli Genome: A Markov Chain Analysis. *Nucleic Acids Research.* 15:2611-2626.

Prum, B., Rodolphe, F., and Turckheim, E. 1995. Finding Words with Unexpected Frequencies in Deoxyribonucleic Acid Sequences. *Journal of the Royal Statistical Society, Ser. B.* 57:205–220.

Reiss, R.-D. 1989. Approximate Distributions of Order Statistics. New York: Springer-Verlag.

Robert, H. F., Suzanne, W.F., and Edward, H. W. 1982. Clinical Epidemiology. London: Williams and Wilkins.

Rogers, L.C., and Pitman, J.W. 1981. Markov Functions. *Ann. Prob.* 9:573–582.

Rouault, A. 1978. Lois de Zipf et Sources Markoviennes. *Annales de l'Institut Henri Poincare-Section B.* 14:169–188.

Rubino, G., and Sericola, B. 1989. On Weak Lumpability in Markov Chain, *J. Appl. Probab.* 26:446–457.

Rubino, G., and Sericola, B. 1991. A Finate Characteristic of Weak Lumpable Markov Process. Part I: The Discrete Time Case. *Stochastic Processes and Their Applications.* 38:195–204.

Schwarz, G. 1978. Estimating the Dimension of a Model. *Annals of Statistics.* 6:461–464.

Simon, H.A. 1955. On a Class of Skew Distribution Functions. *Biometrika.* 42:425–440.

Snyder, E.E. 1994. Identification of Protein Coding Regions in Genomic DNA. Ph.D. Disseration. University of Colorado, Boulder.

Snyder, E.E., and Stormo, G.D. 1997. Identifying Genes in Genomic DNA Sequences. In: DNA and Protein Sequence Analysis: A Practical Approach., Bishop, M.J., and Rawlings, C.J. (eds.). New York: Oxford University Press.

Staden, R., and McLachlan, A.D. 1982. Codon Preference and Its Use in Identifying Protein Coding Regions in Long DNA Sequences, *Nucleic Acids Research.* 10:141–156.

Thomas, M.U. 1977, Computational Methods for lumping Markov Chains. In: Proceedings of the Statistical Computing Section, ASA, Washington. 364–367.

Thomas, M.U., and Barr, D.R. 1977, An Approximate Test of Markov Chain Lumpability, *Journal of the American Statistical Association.* 72:175–179.

Tong, H. 1975. Determination of the Order of a Markov Chain by Akaike's Information Criterion. *J. Appl. Prob.* 12:488–497.

Troll, G., and Graben, P.B. 1998. Zipf's Law Is Not a Consequense of the Central Limit Theorem. *Physical Review E.* 57:1347–1355.

Uberbacher, E.C., and Mural, R.J. 1991. Locating Protein-Coding Regions in Human DNA Sequences by a Multiple Sensor-Neural Network Approach. *Proc. Natl. Acad. Sci. USA.* 88:11261–11265.

Viharos, L. 1993. Asymptotic Distributions of Linear Combinations of Extreme Values *Acta Sci. Math. (Szeged).* 58:211–231.

Viharos, L. 1995. Limiting Theorems for Linear Combination of Extreme Values with Applications to Inference about the Tail of a Distribution. *Acta. Sci. Math. (Szeged).* 60:761–777.

Waterman, M.S. 1995. Introduction to Computational Biology-Maps, Sequences and Genomes. London: Chapman and Hall.

Xu, Y., Mural, R., Shah, M., and Uberbacher, E. 1994. Recognizing Exons in Genomic Sequence Using GRAIL II. In: Genetic Engineering, Vol.16, Stlow, J.K. (ed.). New York: Plenum Press. 241–253.

Zipf, G.K. 1949. Human Behavior and the Principle of Least Effort. Reading, MA: Addison-Wesley.

BIOGRAPHICAL SKETCH

Kil-Sup Lim was born and grew up in Seoul, Republic of Korea. He was awarded a Bachelor of Economics degree in statistics in 1982 followed by a Master of Economics degree in statistics in 1984, from Yonsei University, Seoul, Republic of Korea.

He served as a second lieutenant in the Korean army from 1984 to 1985. After an honorable discharge, he was a researcher at the Korea Institute for Defense Analysis, which has supported his graduate study in the Department of Statistics at the University of Florida. After finishing his study, he plans to return to the institute and continue to serve as a researcher.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Mark C.K. Yang, Chairman
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Richard L. Scheaffer
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Randy L. Carter
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Susan P. McGorray
Research Assistant Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Li-Min Fu
Associate Professor of Computer and Information Science and Engineering

This dissertation was submitted to the Graduate Faculty of the Department of Statistics in the College of Liberal Arts and Sciences and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 1998

_____

Dean, Graduate School